LLM Evaluation

COMP4901Y

Binhang Yuan

Overview

RELAXED
SYSTEM LAB

- A principled benchmark is essential to understand the capacity of LLMs.
- We will discuss:
  - The quality evaluation format and metrics.
  - Concrete quality benchmark examples:
    - General-purpose benchmarks.
    - Coding benchmarks.
    - Math benchmarks.
  - LLM service system evaluation.

LLM Evaluation Metric for Quality

Quality Evaluation Format and Metric

RELAXED SYSTEM LAB

- **Multiple-Choice Question (MCQ)**:
  - Models answer fixed-option questions (A/B/C/D, etc.), selecting one correct choice.
  - Evaluation metric: ***accuracy***, i.e. , the percentage of questions answered correctly.
- **Coding Problem**:
  - Models generate code to solve programming problems. The prompt might be a function description or a problem statement, and the model outputs code.
  - Execute the generated code against the tests.
  - Evaluation metric: a common metric is ***pass@k*** – e.g., pass@1 is the percentage of problems solved on the first attempt, and pass@k indicates if at least one of k tries passes all tests.

Evaluation Format and Metric

RELAXED SYSTEM LAB

- **Open-ended Question**:
  - Evaluation method 1- exact match: a stricter form of ***accuracy*** commonly used in QA evaluation. EM requires the model's output to match the gold answer ***exactly***, down to every character. If there's any deviation (missing a word, extra punctuation, etc.), it counts as incorrect
  - Evaluation method 2 – partial match: compute the precision, recall, and f1 score between the *tokens* in the gold answer and the generated answer:
    - ***Precision:*** the percentage of the answer that is actually correct;
    - ***Recall*** is the percentage of the gold answer that the model captured.
    - ***F1 Score***: The harmonic mean of precision and recall, F1 = 2·precision·recall / (precision + recall).

Partial Match Example

RELAXED
SYSTEM LAB

- Question: "What are the symptoms of influenza?"
  - Gold Answer: "Fever, cough, sore throat, muscle aches, fatigue."
  - LLM Response: "Fever, cough, sore throat, fatigue."
- Tokenization:
  - Reference Tokens: ["fever", "cough", "sore", "throat", "muscle", "aches", "fatigue"]
  - LLM Response Tokens: ["fever", "cough", "sore", "throat", "fatigue"]
  - Correctly Matched Tokens: ["fever", "cough", "sore", "throat", "fatigue"] (Total Matched: 5)
- Calculating Metrics:
  - Precision = Matched Tokens / Total Tokens in LLM Response = 5 / 5 = 1.0
  - Recall = Matched Tokens / Total Tokens in Reference Answer = 5 / 7 ≈ 0.714
  - F1 Score = 2 * (Precision * Recall) / (Precision + Recall) ≈ 0.833

Evaluation Format and Metric

RELAXED SYSTEM LAB

- **Open-ended Question**:
  - Evaluation method 3 - human evaluation: human judgment is the ultimate metric -- but it is usually very hard to scale. Format includes:
    - ***Likert scale ratings***: e.g., rating an output's quality from 1–5;
    - ***Pairwise comparisons:*** show two model outputs and ask which is better.
  - Evaluation method 4 – LLM-as-a-Judge: Have an AI model evaluate outputs to approximate human judgment at scale;
    - This approach should be validated to ensure it aligns with true human preferences.
    - Can generate a similar format to human subjective evaluation.

Human Evaluation

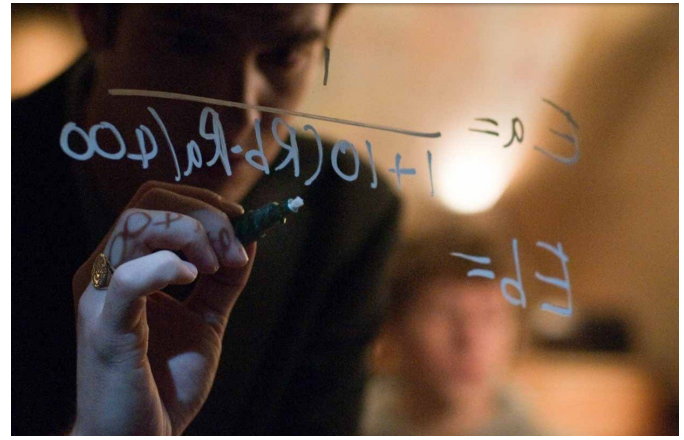
RELAXED
SYSTEM LAB

- Human evaluation can capture more complicated qualities like helpfulness, coherence, creativity, and safety, e.g:
  - Factual accuracy in open-ended answers;
  - Non-deterministic correctness (like whether an explanation is convincing);
  - User preference (which response they found more useful).
  - Issues like a model being too verbose or too terse – things hard to quantify automatically.
- **Pairwise comparisons**: A powerful strategy is to show a human two model outputs for the same prompt and ask which is better (or if they are equal):
  - Easier for people (choosing A vs B);
  - More consistent.
- ***Repeated pairwise battles can produce a ranking of models***.

Human Evaluation — Chatbot Arena

RELAXED SYSTEM LAB

- Basic idea: ELO Rating System:
  - Performance is not measured absolutely; it is inferred from wins, losses, and draws against other players.
  - A player's rating, i.e., expected score R, is their probability of winning plus half their probability of drawing.
  - The rating goes up or down based on wins/losses/draws in these pairwise comparisons.



The Social Network

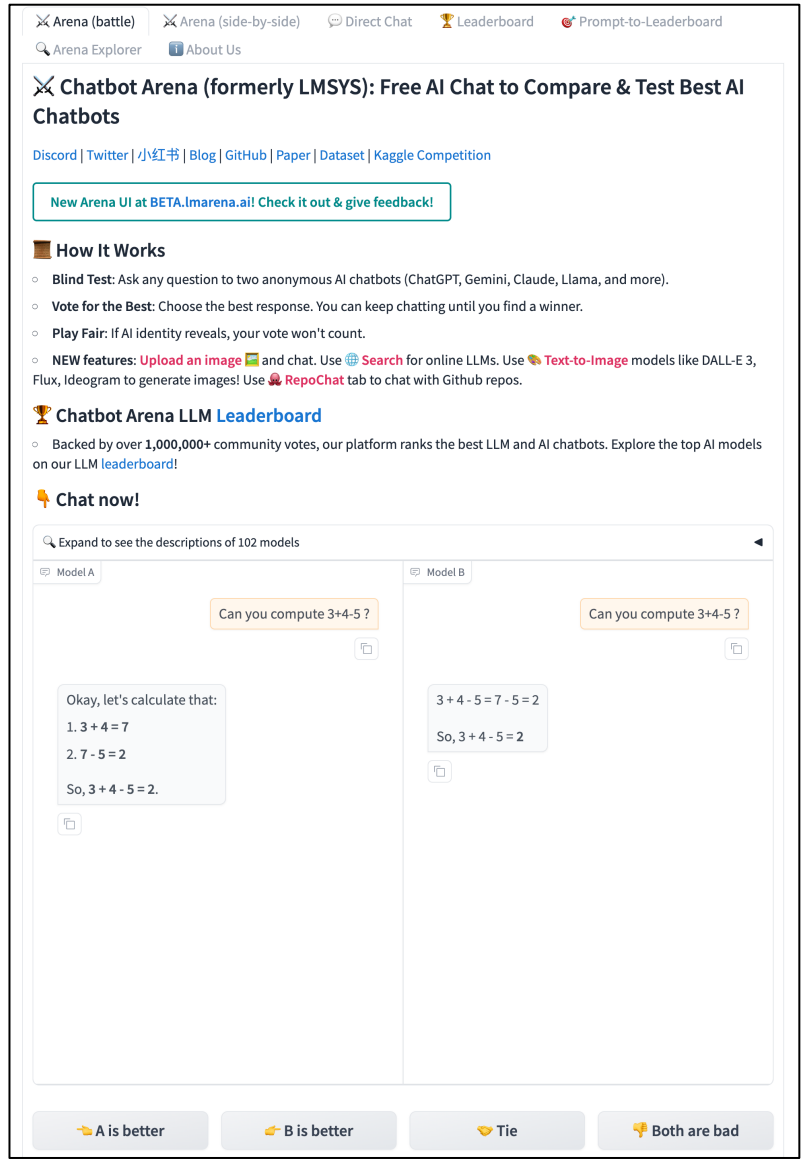
- <u>Expected score calculation:</u>
  - Suppose there are two players, A and B, with ratings R_A and R_B.
  - The system predicts the expected outcome of a match between them, e.g., the expected score for Player A against Player B is calculated using the formula: $E_A = \frac{1}{1+10^{(R_A - R_B)/400}}$

- <u>Rating update after a match:</u>
  - After the match, Player A's rating is updated based on the actual outcome compared to the expected score: $R'_A = R_A + K(S_A - E_A)$.
    - R'_A is the new rating for Player A;
    - K is the development coefficient (commonly 10, 20, or 40), determining the sensitivity of rating changes;
    - S_A is the actual score achieved by Player A (1 for a win, 0.5 for a draw, 0 for a loss)

Human Evaluation — Chatbot Arena

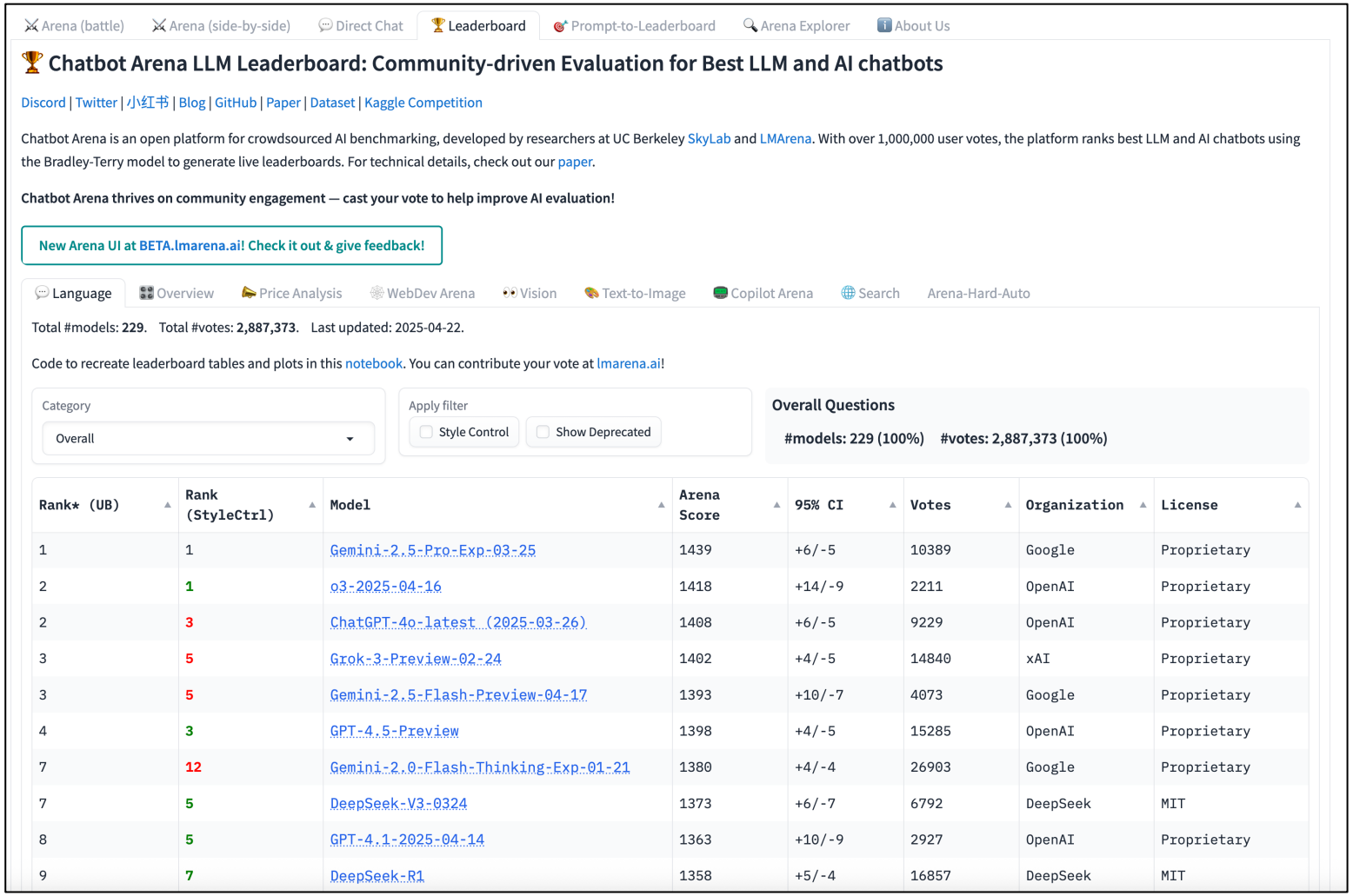
RELAXED SYSTEM LAB

- Chatbot Arena: allows users to chat with two anonymous models side-by-side and then vote which response is better (or declare a tie).
- This generates large-scale pairwise comparison data in a crowd-sourced manner.
- Users don't know which model is which (they are usually labeled Model A and Model B randomly).
- This prevents brand bias (e.g., "I'll pick GPT-4 because I know it's GPT-4") and focuses on the content quality.



Chatbot Arena Leaderboard

RELAXED SYSTEM LAB



⚔ Arena (battle) | ⚔ Arena (side-by-side) | 💬 Direct Chat | 🏆 Leaderboard | 🎯 Prompt-to-Leaderboard | 🔍 Arena Explorer | ℹ️ About Us

🏆 Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots

Discord | Twitter | 小红书 | Blog | GitHub | Paper | Dataset | Kaggle Competition

Chatbot Arena is an open platform for crowdsourced AI benchmarking, developed by researchers at UC Berkeley SkyLab and LMArena. With over 1,000,000 user votes, the platform ranks best LLM and AI chatbots using the Bradley-Terry model to generate live leaderboards. For technical details, check out our paper.

Chatbot Arena thrives on community engagement — cast your vote to help improve AI evaluation!

New Arena UI at BETA.lmarena.ai! Check it out & give feedback!

💬 Language | Overview | Price Analysis | WebDev Arena | 👀 Vision | Text-to-Image | Copilot Arena | 🌐 Search | Arena-Hard-Auto

Total #models: **229**. Total #votes: **2,887,373**. Last updated: 2025-04-22.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at lmarena.ai!

Category: Overall

Apply filter: Style Control | Show Deprecated

Overall Questions

#models: 229 (100%) #votes: 2,887,373 (100%)

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1439	+6/-5	10389	Google	Proprietary
2	1	o3-2025-04-16	1418	+14/-9	2211	OpenAI	Proprietary
2	3	ChatGPT-4o-latest (2025-03-26)	1408	+6/-5	9229	OpenAI	Proprietary
3	5	Grok-3-Preview-02-24	1402	+4/-5	14840	xAI	Proprietary
3	5	Gemini-2.5-Flash-Preview-04-17	1393	+10/-7	4073	Google	Proprietary
4	3	GPT-4.5-Preview	1398	+4/-5	15285	OpenAI	Proprietary
7	12	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	+4/-4	26903	Google	Proprietary
7	5	DeepSeek-V3-0324	1373	+6/-7	6792	DeepSeek	MIT
8	5	GPT-4.1-2025-04-14	1363	+10/-9	2927	OpenAI	Proprietary
9	7	DeepSeek-R1	1358	+5/-4	16857	DeepSeek	MIT

https://lmarena.ai/?leaderboard

(Checked in 2025-04-28)

LLM-as-a-Judge

RELAXED SYSTEM LAB

- Instead of humans, the LLM judge scores each response (for relevance, correctness, etc.), giving a quantitative evaluation.
- This "AI judging AI" approach can quickly evaluate open-ended outputs *at scale*, though care is needed to ensure the scores align with human quality judgments.
- **Pairwise comparison**: The judge LLM is presented with two responses to the same prompt and asked to determine which one is better.
  - Example Prompt: "Given the following two responses to the prompt, which one is more accurate and informative?"
- **Single output scoring (reference-free):** The judge LLM evaluates a single response based on predefined criteria without a reference answer.
  - Example Prompt: "Rate the relevance of the following response on a scale from 1 to 5."
- **Single output scoring (reference-based)**: The judge LLM compares a generated response to a reference answer, evaluating similarity or correctness.
  - Example Prompt: "Compare the generated response to the reference answer and rate its accuracy on a scale from 1 to 5."

LLM-as-a-Judge Example

RELAXED SYSTEM LAB

Question: *"What are the health benefits of regular exercise?"*

Model Responses:

- **Model A:** *"Regular exercise improves cardiovascular health, strengthens muscles, enhances flexibility, and boosts mental well-being."*
- **Model B:** *"Engaging in regular physical activity can lead to better heart health, improved mood, weight management, and increased energy levels."*

Prompt of LLM-as-a-Judge:

You are an expert evaluator. Given a question and two responses from different models, determine which response is more helpful and accurate. Provide a brief explanation for your choice.

Question:

"What are the health benefits of regular exercise?"

Response from Model A:

"Regular exercise improves cardiovascular health, strengthens muscles, enhances flexibility, and boosts mental well-being."

Response from Model B:

"Engaging in regular physical activity can lead to better heart health, improved mood, weight management, and increased energy levels."

Which response is better? Please answer with 'Model A' or 'Model B' and provide a short justification.

LLM-as-a-Judge Response:

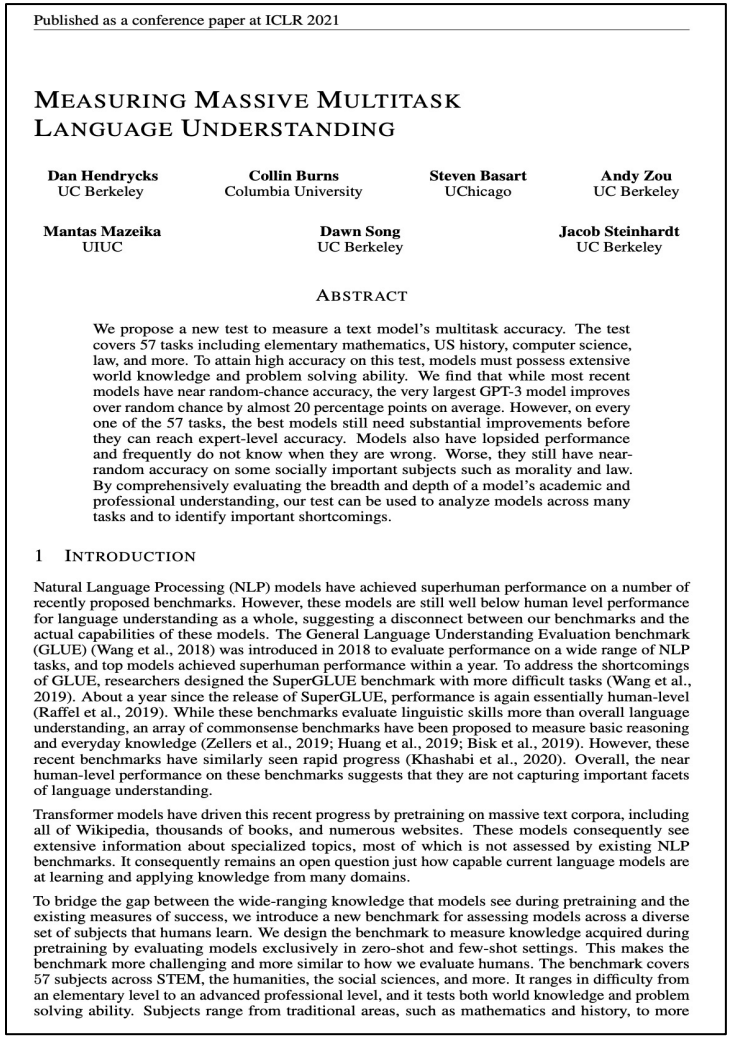
Model A. Justification: Both responses are informative, but Model A provides a more comprehensive overview by mentioning flexibility and mental well-being, which are significant benefits of regular exercise.

General Purpose Benchmark

MMLU

RELAXED SYSTEM LAB

- Massive Multitask Language Understanding (MMLU) is a broad knowledge test that covers 57 subjects spanning STEM, humanities, social sciences, etc., from real examinations (college, high school, professional exams) to evaluate a model's breadth of knowledge.
- It was introduced to challenge models beyond simple language tasks.
- Question set: 15,908 multiple-choice questions (4 options each).
- Exact-match evaluation: Each question has one correct answer. Models are typically tested under zero-shot or few-shot settings and evaluated by accuracy (% of questions answered correctly). Only an exactly correct choice counts.



Published as a conference paper at ICLR 2021

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

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

ABSTRACT

We propose a new test to measure a text model's multitask accuracy. The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more. To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability. We find that while most recent models have near random-chance accuracy, the very largest GPT-3 model improves over random chance by almost 20 percentage points on average. However, on every one of the 57 tasks, the best models still need substantial improvements before they can reach expert-level accuracy. Models also have lopsided performance and frequently do not know when they are wrong. Worse, they still have near-random accuracy on some socially important subjects such as morality and law. By comprehensively evaluating the breadth and depth of a model's academic and professional understanding, our test can be used to analyze models across many tasks and to identify important shortcomings.

1 INTRODUCTION

Natural Language Processing (NLP) models have achieved superhuman performance on a number of recently proposed benchmarks. However, these models are still well below human level performance for language understanding as a whole, suggesting a disconnect between our benchmarks and the actual capabilities of these models. The General Language Understanding Evaluation benchmark (GLUE) (Wang et al., 2018) was introduced in 2018 to evaluate performance on a wide range of NLP tasks, and top models achieved superhuman performance within a year. To address the shortcomings of GLUE, researchers designed the SuperGLUE benchmark with more difficult tasks (Wang et al., 2019). About a year since the release of SuperGLUE, performance is again essentially human-level (Raffel et al., 2019). While these benchmarks evaluate linguistic skills more than overall language understanding, an array of commonsense benchmarks have been proposed to measure basic reasoning and everyday knowledge (Zellers et al., 2019; Huang et al., 2019; Bisk et al., 2019). However, these recent benchmarks have similarly seen rapid progress (Khashabi et al., 2020). Overall, the near human-level performance on these benchmarks suggests that they are not capturing important facets of language understanding.

Transformer models have driven this recent progress by pretraining on massive text corpora, including all of Wikipedia, thousands of books, and numerous websites. These models consequently see extensive information about specialized topics, most of which is not assessed by existing NLP benchmarks. It consequently remains an open question just how capable current language models are at learning and applying knowledge from many domains.

To bridge the gap between the wide-ranging knowledge that models see during pretraining and the existing measures of success, we introduce a new benchmark for assessing models across a diverse set of subjects that humans learn. We design the benchmark to measure knowledge acquired during pretraining by evaluating models exclusively in zero-shot and few-shot settings. This makes the benchmark more challenging and more similar to how we evaluate humans. The benchmark covers 57 subjects across STEM, the humanities, the social sciences, and more. It ranges in difficulty from an elementary level to an advanced professional level, and it tests both world knowledge and problem solving ability. Subjects range from traditional areas, such as mathematics and history, to more

MMLU Improved Versions

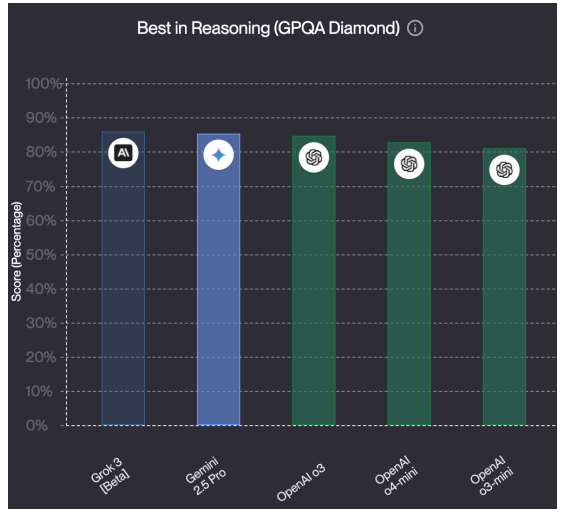
RELAXED SYSTEM LAB

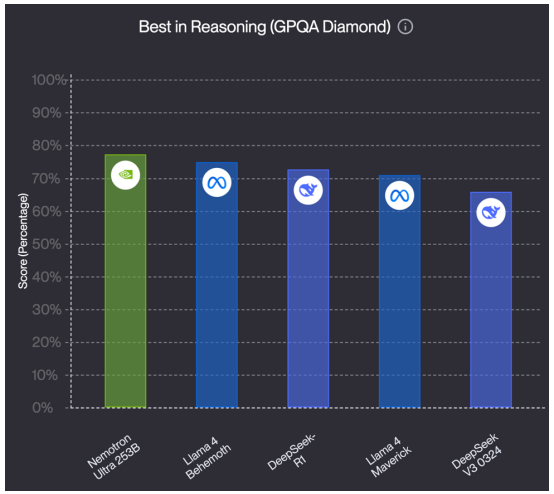
- MMLU-Pro: an enhanced benchmark introduced in 2024 to be more challenging and robust than the original MMLU:
  - Add more challenging, reasoning-focused questions;
  - Each question in MMLU-Pro has 10 answer options instead of 4.
- MMMLU: Multilingual MMLU, an extension of the MMLU benchmark translated into various languages.
  - MMMLU test set has been professionally translated into 14 languages (Chinese, Arabic, Spanish, etc.)
  - checks if a model's broad knowledge holds up in languages other than English: The translated questions maintain the same content difficulty, just expressed in the target language.
  - Same multiple-choice format (4 options) for each question.

GPQA

RELAXED SYSTEM LAB

- GPQA "Google-proof" question answer: a benchmark of challenging multiple-choice questions in biology, physics, and chemistry, written at graduate level.
  - Aims to assess advanced reasoning in scientific domains.
  - The questions are designed such that you can't easily find the exact answer via a quick web search.
  - They often require combining knowledge or understanding an explanation —simulating what a grad student should know, not just copying and pasting from Wikipedia.
- Question set: GPQA consists of 448 multiple-choice questions (with 4 options each) curated by domain experts (e.g., PhD students or professors)
- GPQA Diamond is a high-quality subset of the GPQA benchmark, consisting of the most challenging 198 questions.





https://www.vellum.ai/open-llm-leaderboard

RELAXED
SYSTEM LAB

Coding Benchmark

CodeContests

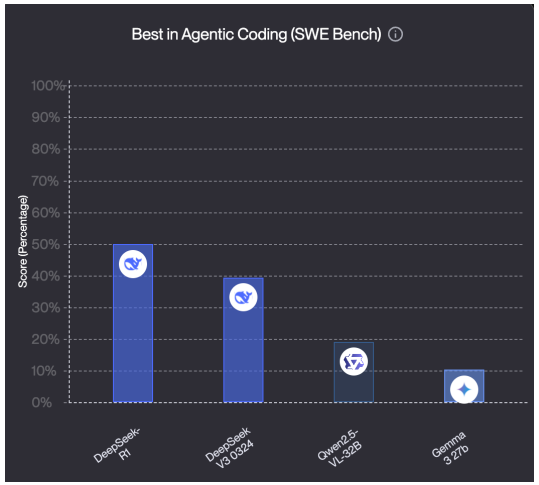
RELAXED
SYSTEM LAB

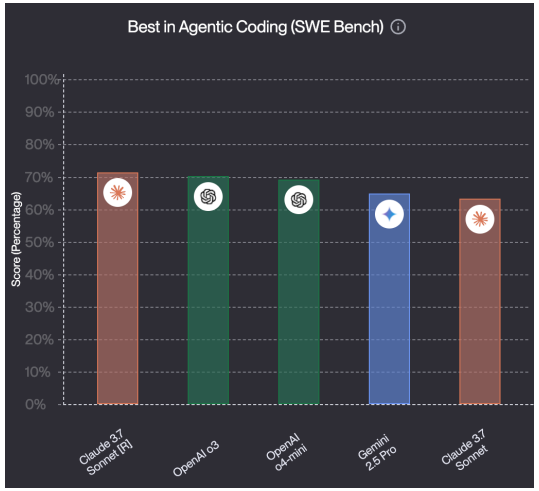
- CodeContests is a benchmark dataset developed to evaluate the capabilities of large language models (LLMs) in solving competitive programming problems.
- The dataset aggregates problems from multiple competitive programming platforms, where each test includes:
  - Problem Statement: A textual description outlining the programming challenge.
  - Input and Output Specifications: Details on the expected input format and the desired output.
  - Public Test Cases: Sample inputs and outputs to illustrate the problem.
- The primary metric for evaluating model performance on CodeContests is pass@k, which measures the percentage of problems solved by at least one of the top-k generated solutions.

SWE-Bench

RELAXED SYSTEM LAB

- SWE-Bench: a benchmark designed to evaluate the ability of large language models (LLMs) to autonomously resolve real-world software engineering issues.
- Each task provides:
  - Issue description: A textual description of a real-world software issue sourced from GitHub.
  - Codebase access: The complete codebase of the repository where the issue exists.
- Evaluation:
  - Ask the model to return the generated patch, using unix's patch program to the codebase. (similar to your GitHub commit)
  - Execute the unit and system tests associated with the task instance.
  - If the patch applies successfully and all of these tests pass we consider the proposed solution to have successfully resolved the issue.





https://www.vellum.ai/open-llm-leaderboard

RELAXED
SYSTEM LAB

Math Benchmark

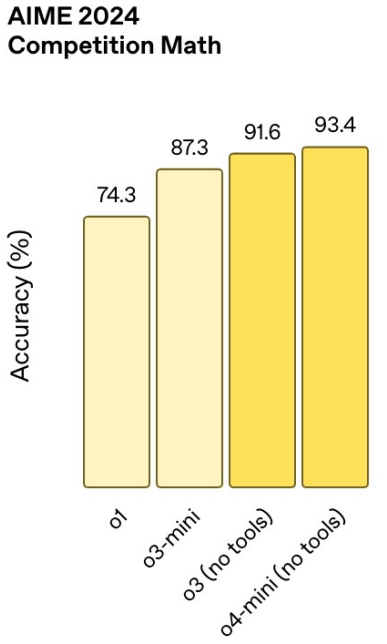
Math-500

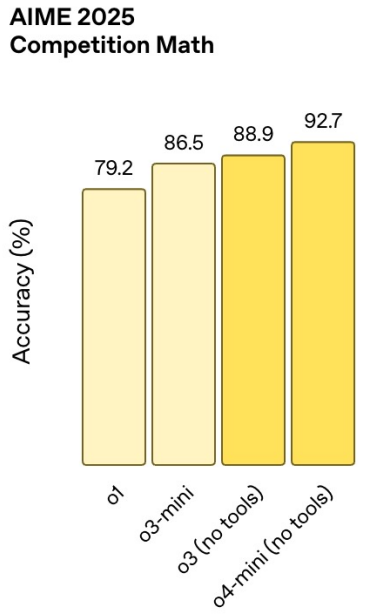
RELAXED
SYSTEM LAB

- Mixed Topic Math (MATH-500): Mixed-difficulty math set, a set of 500 math problems compiled to test a range of mathematical reasoning skills.
- These 500 problems probably cover algebra, geometry, number theory, combinatorics, calculus, etc. They serve as a broad, balanced math benchmark.
- Evaluation format:
  - **Input**: Each problem is given in text form. They range from short questions ("Solve for x: …") to paragraph-long word problems. There's no multiple choice; the model must work it out.
  - **Output**: A simplified exact answer (integer, fraction, expression). The evaluation expects that exact string.
- Evaluation: The metric is Exact Match (EM) — the model's final answer must exactly match the correct answer (usually a simplified expression or number). This again requires full solution correctness.
  - DeepSeek-V3 achieved **90.2% EM** on MATH-500.

AIME

- Advanced math challenge: e.g., AIME 2024 refers to problems from the American Invitational Mathematics Examination 2024.
- AIME is a prestigious high school math contest with very challenging problems – harder than typical school exams, requiring creative problem solving.
  - These problems are designed to require reasoning and multiple steps.
  - They often combine clever insights, algebraic manipulation, number theory, or geometry.
- Format:
  - **Input:** Each problem is given as a word problem (often a paragraph with some math setup).
  - **Output:** The model should output a number (0-999).
- Exact match: AIME problems have an integer answer 0-999. The model must carry through the entire solution correctly to get that exact answer.
  - Latest results from OpenAI reasoning models …





RELAXED
SYSTEM LAB

System Benchmark

LLM API

RELAXED SYSTEM LAB

- LLM API is an important component for you to build you LLM applications.
- **Simplified Integration**: LLM APIs offer a straightforward way to incorporate language understanding into applications. Developers can send text inputs to the API and receive generated responses
- **Scalability**: Using APIs allows applications to scale their language processing capabilities based on demand.

python Copy Edit

```python
import openai

# Replace 'your-api-key' with your actual OpenAI API key
openai.api_key = 'your-api-key'

response = openai.ChatCompletion.create(
    model="gpt-4",
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Can you explain how the OpenAI API works?"}
    ]
)

print(response['choices'][0]['message']['content'])
```

OpenAI API Call Python Example

Evaluating LLM API Service

RELAXED SYSTEM LAB

- When selecting a Large Language Model (LLM) API, besides output quality, it's crucial to consider various system-level factors:
- **Output speed** (tokens per second): This measures how quickly the model generates tokens. Higher output speeds are beneficial for applications requiring rapid responses.
- **Latency** (time to first token): Lower latency ensures faster initial responses, which is critical for real-time applications.
- **End-to-end response time**: This encompasses the total time from sending a request to receiving the complete response. It's a comprehensive metric for assessing user experience.
- **Pricing**: LLM APIs often charge based on the number of input and output tokens processed. It's essential to compare input and output token costs across providers.

API Leaderboard

RELAXED SYSTEM LAB

		FEATURES	MODEL INTELLIGENCE	PRICE	OUTPUT TOKENS/S	LATENCY	END-TO-END RESPONSE TIME		
API PROVIDER	MODEL	CONTEXT WINDOW	ARTIFICIAL ANALYSIS INTELLIGENCE INDEX	BLENDED *USD/1M Tokens*	MEDIAN *Tokens/s*	MEDIAN *First Chunk (s)*	TOTAL *Response (s)*	REASONING *Time (s)*	FURTHER ANALYSIS
OpenAI	o4-mini (high)	200k	70	$1.93	126.6	36.00	39.94	N/A	Model Providers
Microsoft Azure	o4-mini (high)	200k	70	$1.93	79.2	65.21	71.52	N/A	Model Providers
Google	Gemini 2.5 Pro	1m	68	$3.44	214.6	29.55	31.88	N/A	Model Providers
Microsoft Azure	o3	128k	67	$17.50	81.0	38.07	44.25	N/A	Model Providers
xAI	Grok 3 mini Reasoning (high)	131k	67	$0.35	202.0	0.38	12.76	9.90	Model Providers
xAI	Grok 3 mini Reasoning (high) Fast	131k	67	$1.45	214.9	0.46	12.10	9.31	Model Providers
OpenAI	o3-mini (high)	200k	66	$1.93	156.9	46.97	50.16	N/A	Model Providers
Microsoft Azure	o3-mini (high)	200k	66	$1.93	205.5	40.61	43.05	N/A	Model Providers
OpenAI	o3-mini	200k	63	$1.93	156.3	14.34	17.54	N/A	Model Providers
Microsoft Azure	o3-mini	200k	63	$1.93	201.4	12.64	15.13	N/A	Model Providers
OpenAI	o1	200k	62	$26.25	68.9	50.87	58.13	N/A	Model Providers
Microsoft Azure	o1	200k	62	$26.25	114.9	26.30	30.65	N/A	Model Providers
Google	Gemini 2.5 Flash (Reasoning) (AI_Studio)	1m	60	$0.99	347.5	8.35	9.79	N/A	Model Providers
Lambda	DeepSeek R1	164k	60	$0.95	38.2	0.55	75.05	61.41	Model Providers
deepseek	DeepSeek R1	64k	60	$0.96	23.5	3.86	125.12	99.96	Model Providers
Hyperbolic	DeepSeek R1	128k	60	$2.00	71.5	2.11	41.93	32.83	Model Providers
aws	DeepSeek R1	128k	60	$2.36	65.2	0.44	44.08	35.97	Model Providers

https://artificialanalysis.ai/leaderboards/providers

Reference

RELAXED
SYSTEM LAB

- https://docs.mistral.ai/guides/evaluation/
- https://medium.com/@saveriomazza/arena-elo-rating-system-5655e16fead5
- https://colab.research.google.com/drive/1KdwokPjirkTmpO_P1WByFNFiqxWQquwH
- https://lmarena.ai/?leaderboard
- https://www.evidentlyai.com/llm-guide/llm-as-a-judge
- https://en.wikipedia.org/wiki/MMLU
- https://arxiv.org/pdf/2406.01574
- https://huggingface.co/datasets/openai/MMMLU
- https://arxiv.org/pdf/2311.12022
- https://arxiv.org/abs/2310.06770
- https://huggingface.co/datasets/HuggingFaceH4/MATH-500
- https://www.vals.ai/benchmarks/aime-2025-03-11