# 5. Evaluation

Hyunseung Cha

# Benchmarks

| | GPT-4<br>Evaluated few-shot | GPT-3.5<br>Evaluated few-shot | LM SOTA<br>Best external LM evaluated few-shot | SOTA<br>Best external model (incl. benchmark-specific tuning) |
|---|---|---|---|---|
| MMLU [49]<br>Multiple-choice questions in 57 subjects (professional & academic) | **86.4%**<br>5-shot | 70.0%<br>5-shot | 70.7%<br>5-shot U-PaLM [50] | 75.2%<br>5-shot Flan-PaLM [51] |
| HellaSwag [52]<br>Commonsense reasoning around everyday events | **95.3%**<br>10-shot | 85.5%<br>10-shot | 84.2%<br>LLaMA (validation set) [28] | 85.6<br>ALUM [53] |
| AI2 Reasoning Challenge (ARC) [54]<br>Grade-school multiple choice science questions. Challenge-set. | **96.3%**<br>25-shot | 85.2%<br>25-shot | 85.2%<br>8-shot PaLM [55] | 86.5%<br>ST-MOE [18] |
| WinoGrande [56]<br>Commonsense reasoning around pronoun resolution | **87.5%**<br>5-shot | 81.6%<br>5-shot | 85.1%<br>5-shot PaLM [3] | 85.1%<br>5-shot PaLM [3] |
| HumanEval [43]<br>Python coding tasks | **67.0%**<br>0-shot | 48.1%<br>0-shot | 26.2%<br>0-shot PaLM [3] | 65.8%<br>CodeT + GPT-3.5 [57] |
| DROP [58] (F1 score)<br>Reading comprehension & arithmetic. | 80.9<br>3-shot | 64.1<br>3-shot | 70.8<br>1-shot PaLM [3] | **88.4**<br>QDGAT [59] |
| GSM-8K [60]<br>Grade-school mathematics questions | **92.0%***<br>5-shot chain-of-thought | 57.1%<br>5-shot | 58.8%<br>8-shot Minerva [61] | 87.3%<br>Chinchilla + SFT+ORM-RL, ORM reranking [62] |

## Astronomy

**What is true for a type-Ia supernova?**

A. This type occurs in binary systems.
B. This type occurs in young galaxies.
C. This type produces gamma-ray bursts.
D. This type produces high amounts of X-rays.

Answer: A

## High School Biology

**In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of**

A. directional selection.
B. stabilizing selection.
C. sexual selection.
D. disruptive selection

Answer: A

# Collection of benchmarks

## Knowledge and Language Understanding

### Massive Multitask Language Understanding (MMLU)

- **Description:** Measures general knowledge across 57 different subjects, ranging from STEM to social sciences.
- **Purpose:** To assess the LLM's understanding and reasoning in a wide range of subject areas.
- **Relevance:** Ideal for multifaceted AI systems that require extensive world knowledge and problem solving ability.
- **Source:** Measuring Massive Multitask Language Understanding
- **Resources:**
  - MMLU GitHub
  - MMLU Dataset

### AI2 Reasoning Challenge (ARC)

- **Description:** Tests LLMs on grade-school science questions, requiring both deep general knowledge and reasoning abilities.
- **Purpose:** To evaluate the ability to answer complex science questions that require logical reasoning.
- **Relevance:** Useful for educational AI applications, automated tutoring systems, and general knowledge assessments.
- **Source:** Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge
- **Resources:**
  - ARC Dataset: HuggingFace
  - ARC Dataset: Allen Institute

https://github.com/leobeeson/llm_benchmarks

# EleutherAI lm evaluation

https://github.com/EleutherAI/lm-evaluation-harness

실제로 제일 자주 참고하고 쓴 것

# **ROUGE:** Recall-Oriented Understudy Gisting Evaluation

텍스트 요약의 품질 평가 방법.

기계가 생성한 요약을 인간이 작성한 참고 요약과 비교하여 얼마나 유사한지를 측정. ROUGE는 Recall을 중심으로 평가.

1. **ROUGE-N**: n-그램의 일치도를 측정합니다. 여기서 N은 1, 2, 3 등 다양한 값이 될 수 있습니다.

$$\text{ROUGE-N} = \frac{\text{참고 요약의 n-그램 중 기계 요약에 포함된 n-그램 수}}{\text{참고 요약의 n-그램 수}}$$

2. **ROUGE-L**: 장문 공통 서브시퀀스(Longest Common Subsequence, LCS)를 사용하여 측정합니다. LCS는 두 텍스트에서 순서를 유지하면서 가장 긴 공통 부분을 찾는 방법입니다.

$$\text{ROUGE-L} = \frac{LCS(\text{참고 요약}, \text{기계 요약})}{\text{참고 요약의 단어 수}}$$

3. **ROUGE-S**: 빅람(그램)과 스킵-빅람(Skip-Bigram)을 사용합니다. 스킵-빅람은 원래 순서를 유지하면서 단어를 건너뛸 수 있는 빅람입니다.

$$\text{ROUGE-S} = \frac{\text{참고 요약의 스킵-빅람 중 기계 요약에 포함된 스킵-빅람 수}}{\text{참고 요약의 스킵-빅람 수}}$$

https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb

# **ROUGE:** Recall-Oriented Understudy Gisting Evaluation

```python
import evaluate

# Load the ROUGE evaluation metric
rouge = evaluate.load('rouge')

# Define the candidate predictions and reference sentences
predictions = ["hello there", "general kenobi"]
references = ["hello there", "general kenobi"]

# Compute the ROUGE score
results = rouge.compute(predictions=predictions, references=references)

# Print the results
print(results)
```

```
{'rouge1': 1.0, 'rouge2': 1.0, 'rougeL': 1.0, 'rougeLsum': 1.0}
```

https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb

# **BLEU:** Bilingual Evaluation Understudy

기계 번역의 품질을 평가하는 방법.

번역된 문장을 사람의 번역(참고 번역)과 비교하여 얼마나 유사한지를 측정.

주로 번역의 정확성을 객관적으로 평가하기 위해 사용

1. **정밀도(Precision)**: 기계 번역에서 참고 번역과 일치하는 n-그램의 비율입니다.

$$\text{Precision} = \frac{\text{기계 번역의 n-그램 중 일치하는 n-그램 수}}{\text{기계 번역의 n-그램 수}}$$

2. **클립된 정밀도(Clipped Precision)**: 각 n-그램이 참고 번역에서 최대 몇 번 일치할 수 있는지를 제한하여 계산합니다.

   예를 들어, 참고 번역에 "the cat"이 한 번만 등장하면, 기계 번역에 "the cat"이 여러 번 등장해도 한 번만 일치로 계산합니다.

3. **브레버티 페널티(Brevity Penalty, BP)**: 기계 번역이 너무 짧은 경우 페널티를 줍니다.

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases}$$

   여기서 $c$는 기계 번역의 총 단어 수, $r$는 참고 번역의 총 단어 수입니다.

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

# **BLEU:** Bilingual Evaluation Understudy

```
import evaluate

# Define the candidate predictions and reference sentences
predictions = ["hello there general kenobi", "foo bar foobar"]
references = [["hello there general kenobi", "hello there !"],["foo bar foobar"]]

# Load the BLEU evaluation metric
bleu = evaluate.load("bleu")

# Compute the BLEU score
results = bleu.compute(predictions=predictions, references=references)

# Print the results
print(results)
```

```
{'bleu': 1.0,
 'precisions': [1.0, 1.0, 1.0, 1.0],
 'brevity_penalty': 1.0,
 'length_ratio': 1.1666666666666667,
 'translation_length': 7,
 'reference_length': 6}
```

https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb

https://arize.com/blog-course/llm-evaluation-the-definitive-guide/

# Metrics by types

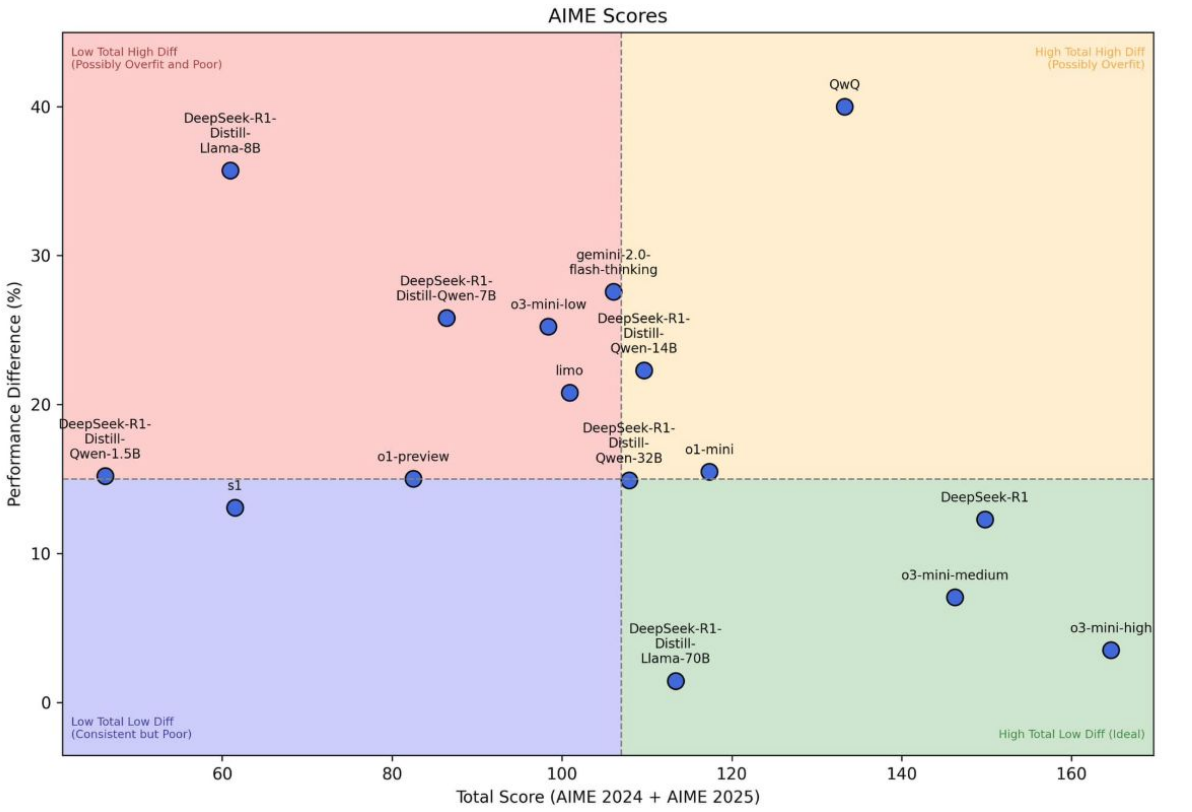
| Type | Description | Example Metrics |
|---|---|---|
| Diversity | Examines the versatility of foundation models in responding to different types of queries | Fluency, Perplexity, ROUGE scores |
| User Feedback | Goes beyond accuracy to look at response quality in terms of coherence and usefulness | Coherence, Quality, Relevance |
| Ground Truth-Based Metrics | Compares a RAG system's responses to a set of predefined, correct answers | Accuracy, F1 score, Precision, Recall |
| Answer Relevance | How relevant the LLM's response is to a given user's query. | Binary classification (Relevant/Irrelevant) |
| QA Correctness | Based on retrieved data, is an answer to a question correct? | Binary classification (Correct/Incorrect) |
| Hallucinations | Looking at LLM hallucinations with regard to retrieved context | Binary classification (Factual/Hallucinated) |
| Toxicity | Are responses racist, biased, or toxic? | Disparity Analysis, Fairness Scoring, Binary classification (Non-Toxic/Toxic) |

# 문제점

- Benchmark 가 너무 다양하고, Benchmark 점수가 높다고 내가 원하는 성능도 좋다는 보장이 없음
- Abusing 이 심하고, 모델 발전에 따라 Benchmark 점수 인플레이션이 생김 (MMLU 이후 MMLU pro 가 나옴)
- LLM 을 쓸 때, 결과가 좋지 않으면 다시 묻는 등 prompt engineering 이 잘 반영되지 않았음
- 최종 사용자가 사람이고, 정답이 명확하지 않은 태스크는 결국 사람이 평가해야 (조금 더) 확신할 수 있음
- 사람이 일일이 (거의) 모든 케이스에 대해 평가하는 것은 불가능에 가까움

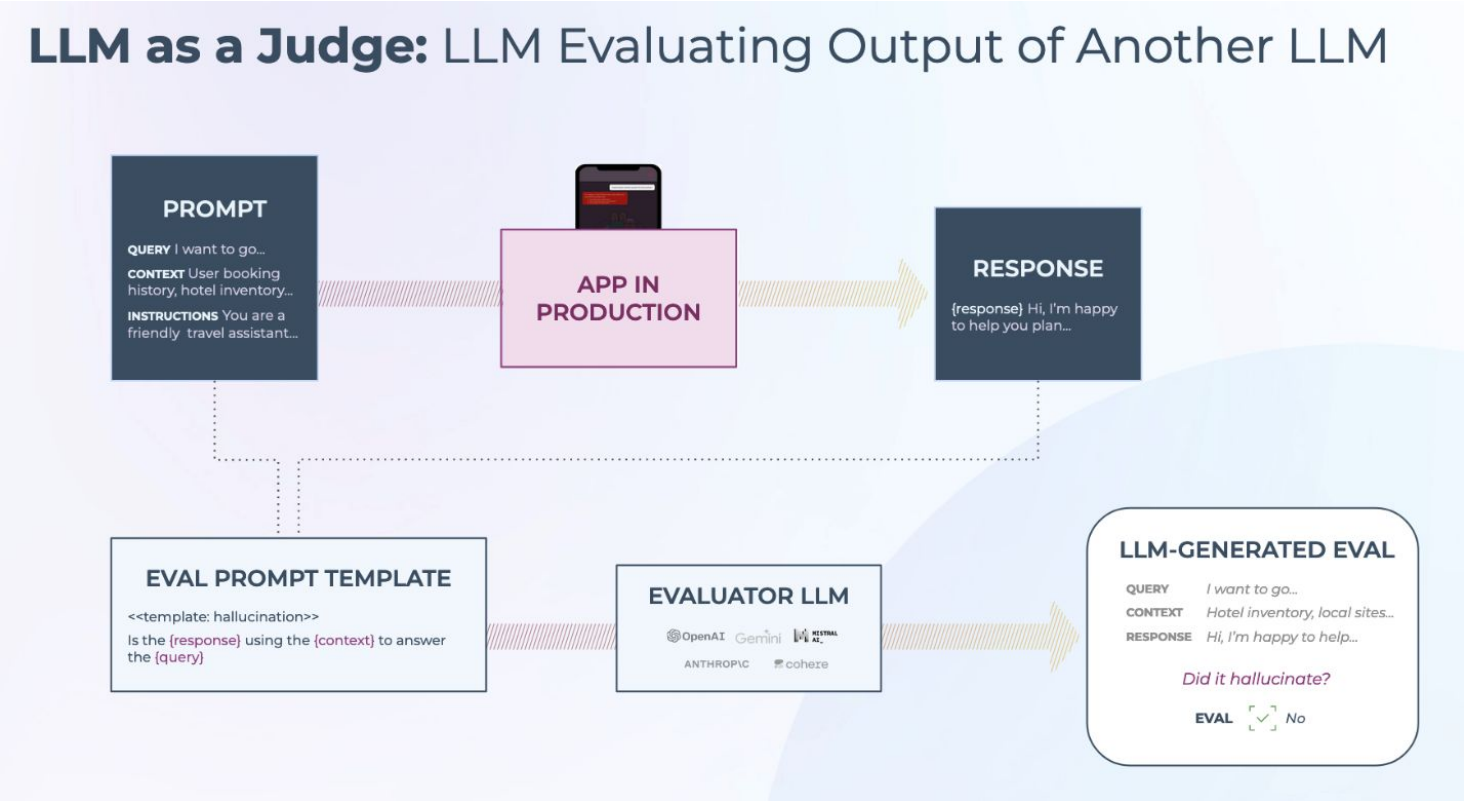
# Abusing?

벤치마크 출처: https://github.com/GAIR-NLP/AIME-Preview



https://arize.com/blog-course/llm-evaluation-the-definitive-guide/

# LLM as a Judge



https://arxiv.org/pdf/2306.05685

https://velog.io/@you-genie/MT-Bench

# MT bench

LLM as a judge 에서 Multi turn 성능을 측정

(LLM 끼리 얼마나 일치하는지, LLM - 사람 간에 얼마나 일치하는지)

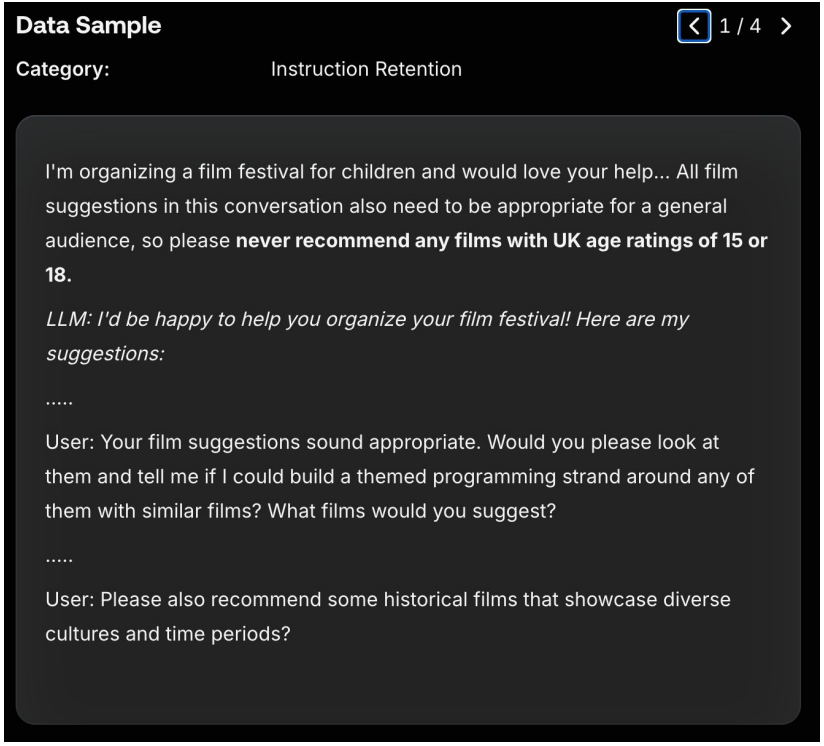
Table 1: Sample multi-turn questions in MT-bench.

| Category | | Sample Questions |
|---|---|---|
| Writing | 1st Turn | Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions. |
| | 2nd Turn | Rewrite your previous response. Start every sentence with the letter A. |
| Math | 1st Turn | Given that $f(x) = 4x^3 - 9x - 14$, find the value of $f(2)$. |
| | 2nd Turn | Find $x$ such that $f(x) = 0$. |
| Knowledge | 1st Turn | Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies ... |
| | 2nd Turn | Now, explain them again like I'm five. |

https://scale.com/leaderboard/multichallenge

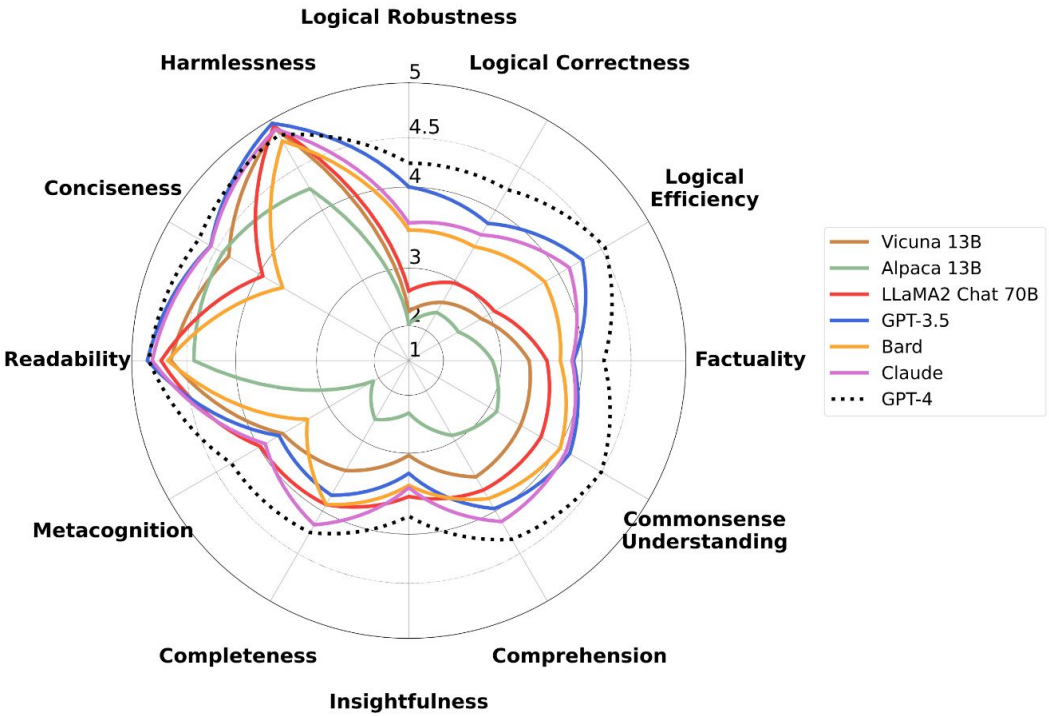
# MultiChallenge

Evaluation for multi turn conversation with user



https://github.com/kaistAI/FLASK

# FLASK: Alignment skill sets




LM 의 언어 능력을 역량별로 나눠서 평가할 수 있도록.

기존 유명 벤치보다 최에 나왔고, 근래 많이 쓰임.

모델 간 입체적으로 비교할 수 있음

https://chat.lmsys.org/?leaderboard

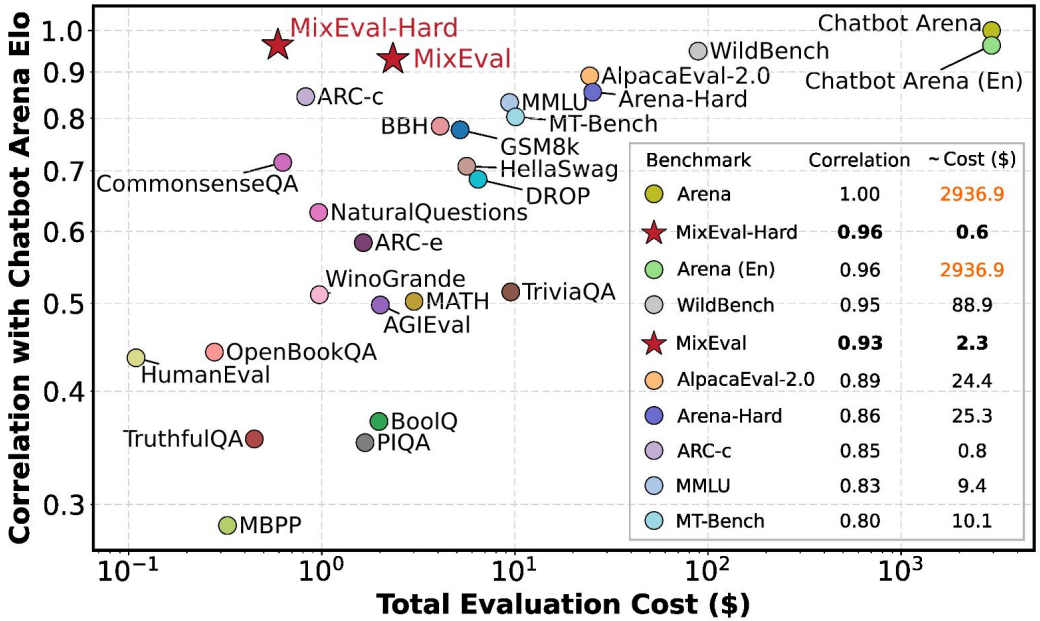
# LMSYS Chatbot Arena Leaderboard

| Rank* (UB) | Rank (StyleCtrl) | Model | Arena Score | 95% CI | Votes | Organization | License |
|---|---|---|---|---|---|---|---|
| 1 | 2 | Grok-3-Preview-02-24 | 1406 | +8/-6 | 9109 | xAI | Proprietary |
| 1 | 1 | GPT-4.5-Preview | 1400 | +5/-6 | 8596 | OpenAI | Proprietary |
| 3 | 6 | Gemini-2.0-Flash-Thinking-Exp-01-21 | 1383 | +6/-4 | 21124 | Google | Proprietary |
| 3 | 3 | Gemini-2.0-Pro-Exp-02-05 | 1380 | +4/-4 | 19038 | Google | Proprietary |
| 3 | 2 | ChatGPT-4o-latest (2025-01-29) | 1375 | +6/-4 | 20936 | OpenAI | Proprietary |
| 6 | 4 | DeepSeek-R1 | 1360 | +7/-5 | 11507 | DeepSeek | MIT |
| 6 | 10 | Gemini-2.0-Flash-001 | 1355 | +4/-5 | 16845 | Google | Proprietary |
| 6 | 3 | o1-2024-12-17 | 1352 | +4/-6 | 23441 | OpenAI | Proprietary |
| 8 | 10 | Gemma-3-27B-it | 1340 | +8/-8 | 5028 | Google | Gemma |
| 9 | 10 | Qwen2.5-Max | 1339 | +4/-5 | 15607 | Alibaba | Proprietary |
| 9 | 7 | o1-preview | 1335 | +4/-4 | 33187 | OpenAI | Proprietary |

+ Routing to the most suitable LLM based on the task can improve the ELO score.

https://mixeval.github.io/
https://huggingface.co/datasets/MixEval/MixEval/viewer

# MixEval: Benchmark mixture




| Benchmark | Correlation | ~Cost ($) |
|---|---|---|
| Arena | 1.00 | 2936.9 |
| MixEval-Hard | **0.96** | **0.6** |
| Arena (En) | 0.96 | 2936.9 |
| WildBench | 0.95 | 88.9 |
| MixEval | **0.93** | **2.3** |
| AlpacaEval-2.0 | 0.89 | 24.4 |
| Arena-Hard | 0.86 | 25.3 |
| ARC-c | 0.85 | 0.8 |
| MMLU | 0.83 | 9.4 |
| MT-Bench | 0.80 | 10.1 |

Benchmark correlations (%) with Chatbot Arena Elo, against the total costs of evaluating a single GPT-3.5-Turbo-0125 model. MixEval and MixEval-Hard show the highest correlations with Arena Elo and Arena Elo (En) among leading benchmarks. We reference the crowdsourcing price for Amazon Mechanical Turk ($0.05 per vote) when estimating the cost of evaluating a single model on Chatbot Arena (approximately $2,936). Chatbot Arena is prohibitively expensive, while MixEval and MixEval-Hard are cheap and cost-effective alternatives. For more details, please refer to our paper.

(상대적으로 신뢰도가 높은) Chatbot Arena 과의 correlation 이 높고, 높은 correlation 을 고려했을 때 다른 벤치 마크 대비 필요 연산량이 작음.

https://huggingface.co/datasets/openai/MMMLU
https://github.com/openai/simple-evals/blob/main/multilingual_mmlu_benchmark_results.md

# Multilingual MMLU (MMMLU)

| Language | o1-preview | gpt-4o-2024-08-06 | o1-mini | gpt-4o-mini-2024-07-18 |
|---|---|---|---|---|
| Arabic | **0.8821** | 0.8155 | **0.7945** | 0.7089 |
| Bengali | **0.8622** | 0.8007 | **0.7725** | 0.6577 |
| Chinese (Simplified) | **0.8800** | 0.8335 | **0.8180** | 0.7305 |
| English (not translated) | **0.9080** | 0.8870 | **0.8520** | 0.8200 |
| French | **0.8861** | 0.8437 | **0.8212** | 0.7659 |
| German | **0.8573** | 0.8292 | **0.8122** | 0.7431 |
| Hindi | **0.8782** | 0.8061 | **0.7887** | 0.6916 |
| Indonesian | **0.8821** | 0.8344 | **0.8174** | 0.7452 |
| Italian | **0.8872** | 0.8435 | **0.8222** | 0.7640 |
| Japanese | **0.8788** | 0.8287 | **0.8129** | 0.7255 |
| Korean | **0.8815** | 0.8262 | **0.8020** | 0.7203 |
| Portuguese (Brazil) | **0.8859** | 0.8427 | **0.8243** | 0.7677 |
| Spanish | **0.8893** | 0.8493 | **0.8303** | 0.7737 |
| Swahili | **0.8479** | 0.7708 | **0.7015** | 0.6191 |
| Yoruba | **0.7373** | 0.6195 | **0.5807** | 0.4583 |

https://github.com/mlfoundations/evalchemy

# Evalchemy

## 📚 Available Tasks

### Built-in Benchmarks

- All tasks from LM Evaluation Harness
- Custom instruction-based tasks (found in `eval/chat_benchmarks/`):
  - **MTBench**: Multi-turn dialogue evaluation benchmark
  - **WildBench**: Real-world task evaluation
  - **RepoBench**: Code understanding and repository-level tasks
  - **MixEval**: Comprehensive evaluation across domains
  - **IFEval**: Instruction following capability evaluation
  - **AlpacaEval**: Instruction following evaluation
  - **HumanEval**: Code generation and problem solving
  - **ZeroEval**: Logical reasoning and problem solving
  - **MBPP**: Python programming benchmark
  - **Arena-Hard-Auto** (Coming soon): Automatic evaluation tool for instruction-tuned LLMs
  - **SWE-Bench** (Coming soon): Evaluating large language models on real-world software issues
  - **SafetyBench** (Coming soon): Evaluating the safety of LLMs
  - **Berkeley Function Calling Leaderboard** (Coming soon): Evaluating ability of LLMs to use APIs

# Hard Datasets

Humanity's Last Exam →
Frontier Multimodal Benchmark

Learn More

| | Model | Accuracy | 95% CI |
|---|---|---|---|
| 1 | Claude 3.7 Sonnet Thinking (Febr... | 8.93 | +1.08 / -1.08 |
| 1 | o1 (December 2024) | 8.81 | +1.07 / -1.07 |
| 1 | Gemini 2.0 Flash Thinking (Janua... | 7.22 | +0.98 / -0.98 |
| 1 | Gemini 2.0 Pro Experimental (Feb... | 7.07 | +0.97 / -0.97 |
| 3 | GPT-4.5 Preview (February 2025) | 6.41 | +0.92 / -0.92 |
| 3 | Llama 3.2 90B Vision Instruct | 5.52 | +0.86 / -0.86 |
| 5 | Gemini-1.5-Pro-002 | 5.22 | +0.84 / -0.84 |
| 5 | Gemini 2.0 Flash Experimental (D... | 5.19 | +0.84 / -0.84 |
| 5 | Gemini 2.0 Flash | 5.07 | +0.83 / -0.83 |
| 5 | Claude 3.7 Sonnet (February 2025) | 5.04 | +0.83 / -0.83 |

View Full Ranking →

https://scale.com/leaderboard

# Recap: Evaluation Process

1. Metrics selection
    a. Retrieval: Context relevance
    b. Generation: Faithfulness (+correctness)
    c. Latency
    d. Toxicity-free
2. Evaluator preparation
    a. Logging: Langsmith, Phoenix
    b. LLM as a judge
    c. Human evaluation
    d. Benchmark dataset
3. Golden dataset (Synthetic data with LLM + Manual tuning)
4. Evaluation & feedback

# OpenAI Evaluation

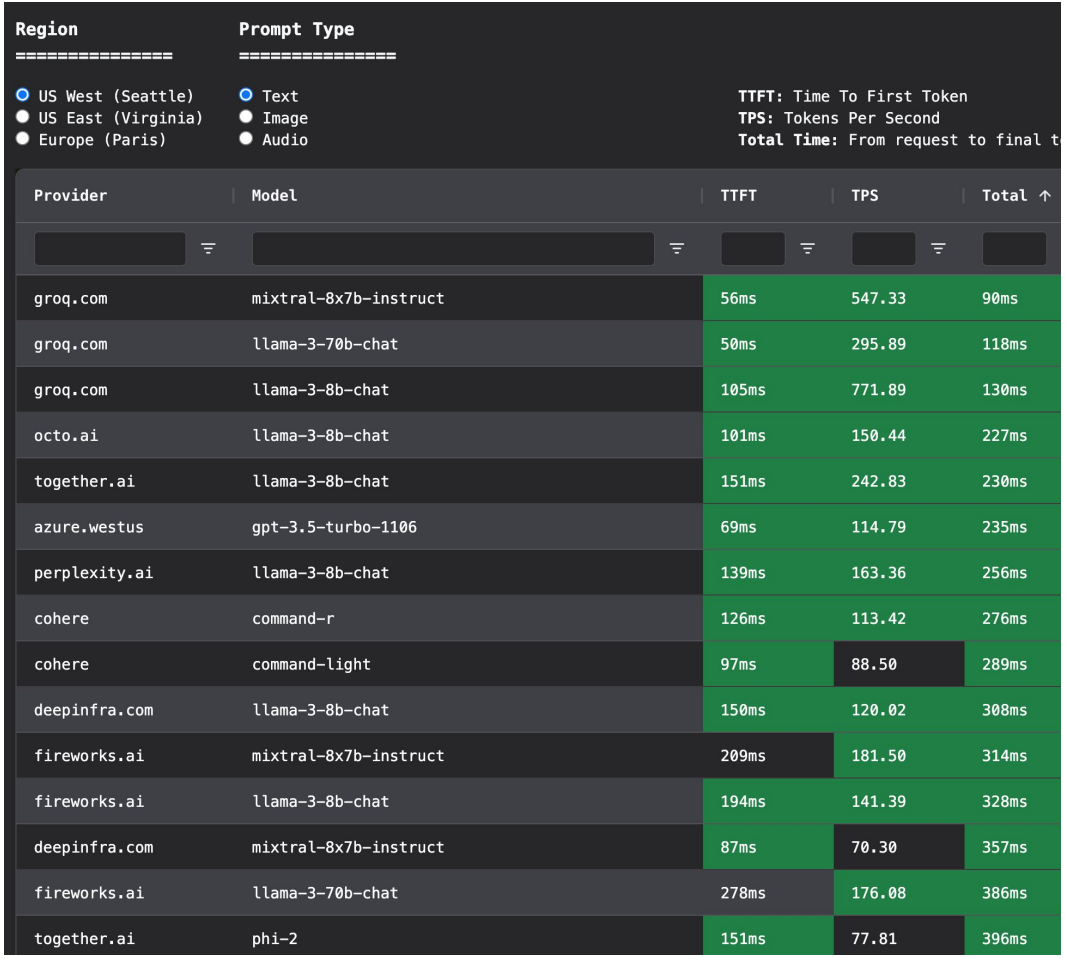
1. Metrics selection
    a. Retrieval: Context relevance
    b. Generation: Faithfulness (+correctness)
    c. Latency
    d. Toxicity-free
2. Evaluator preparation
    a. Logging: Langsmith, Phoenix
    b. LLM as a judge
    c. Human evaluation
    d. Benchmark dataset
3. Golden dataset (Synthetic data with LLM + Manual tuning)
4. Evaluation & feedback

https://thefastest.ai/

# Other metrics

Model + HW + SW

Batch, caching



Region

- (•) US West (Seattle)
- ( ) US East (Virginia)
- ( ) Europe (Paris)

Prompt Type

- (•) Text
- ( ) Image
- ( ) Audio

**TTFT:** Time To First Token
**TPS:** Tokens Per Second
**Total Time:** From request to final t

| Provider | Model | TTFT | TPS | Total ↑ |
|---|---|---|---|---|
| groq.com | mixtral-8x7b-instruct | 56ms | 547.33 | 90ms |
| groq.com | llama-3-70b-chat | 50ms | 295.89 | 118ms |
| groq.com | llama-3-8b-chat | 105ms | 771.89 | 130ms |
| octo.ai | llama-3-8b-chat | 101ms | 150.44 | 227ms |
| together.ai | llama-3-8b-chat | 151ms | 242.83 | 230ms |
| azure.westus | gpt-3.5-turbo-1106 | 69ms | 114.79 | 235ms |
| perplexity.ai | llama-3-8b-chat | 139ms | 163.36 | 256ms |
| cohere | command-r | 126ms | 113.42 | 276ms |
| cohere | command-light | 97ms | 88.50 | 289ms |
| deepinfra.com | llama-3-8b-chat | 150ms | 120.02 | 308ms |
| fireworks.ai | mixtral-8x7b-instruct | 209ms | 181.50 | 314ms |
| fireworks.ai | llama-3-8b-chat | 194ms | 141.39 | 328ms |
| deepinfra.com | mixtral-8x7b-instruct | 87ms | 70.30 | 357ms |
| fireworks.ai | llama-3-70b-chat | 278ms | 176.08 | 386ms |
| together.ai | phi-2 | 151ms | 77.81 | 396ms |

# Trend following? Update?

DeepSeek V3 updated

Gemini pro 2.5

Qwen 2.5 VL

https://uploads-ssl.webflow.com/64e655d42d3be60f582d0472/668fa6d33f471278ecac72ec_Lynx_Hallucination_Model.pdf
https://www.patronus.ai/blog/lynx-state-of-the-art-open-source-hallucination-detection-model

# Optional) LLM to detect hallucination (by lynx)

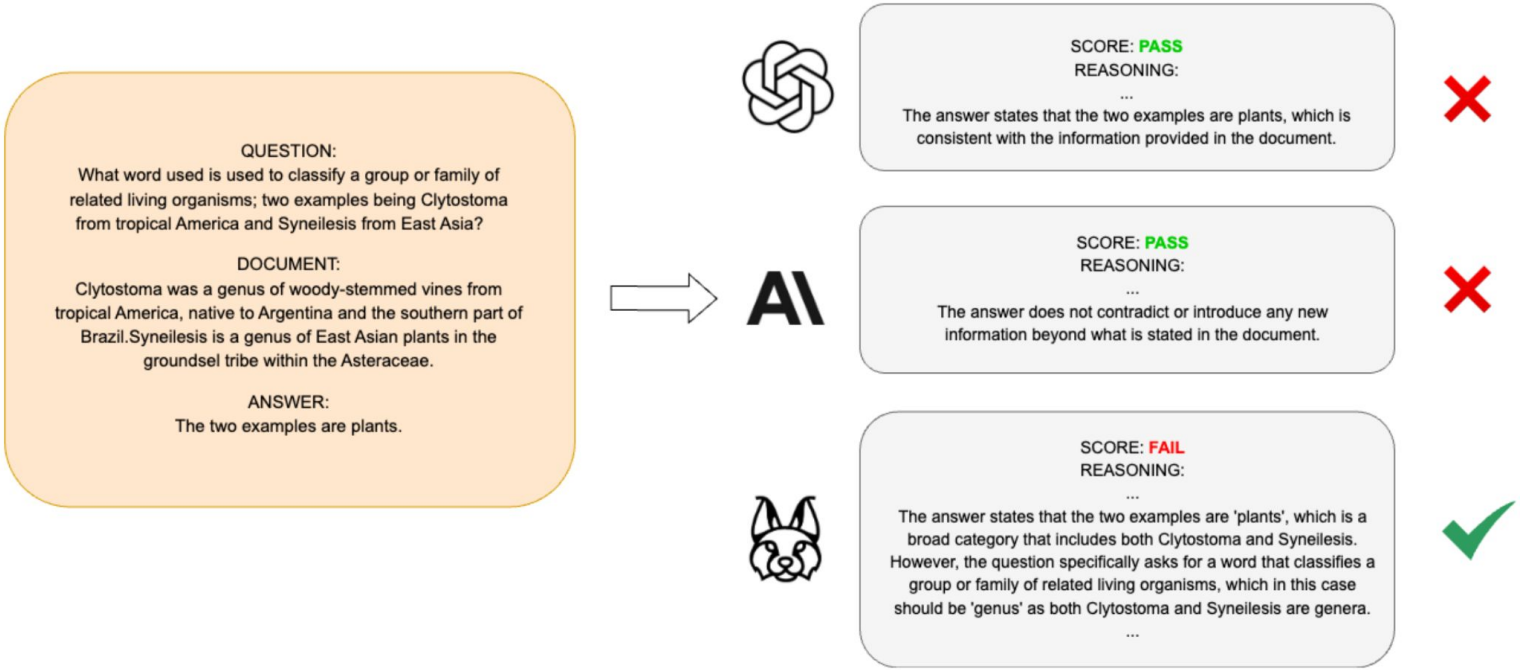



**Figure 1:** LLM-as-a-judge responses of GPT-4o, Claude-3-Sonnet and LYNX (70B) for a Question Answering example from HaluEval.

https://huggingface.co/datasets/PatronusAI/HaluBench

# Optional) HaluBench

Search this dataset

SQL Console

| id<br>string · lengths<br>33↔36 6.7% | passage<br>string · lengths<br>30↔3.23k 91.6% | question<br>string · lengths<br>10↔77 40.2% | answer<br>string · lengths<br>1↔193 88.6% | label<br>string · classes<br>FAIL 48.1% | source_ds<br>string · classes<br>DROP 6.7% |
|---|---|---|---|---|---|
| d3fb4c3c-d21b-480a-baa0-98d6d0d17c1d | Hoping to rebound from the road loss t… | Which team scored the longest field goal… | ['Rams', 'second', 'Marc Bulger', 'Kevin… | FAIL | DROP |
| 8603663e-c53b-46db-a482-a867f12ff3b4 | As of the census of 2000, there were… | How many percent were not Irish? | 87.1 | FAIL | DROP |
| c63a73e5-2c91-489b-bd24-af150ddfa82c | Hoping to rebound from the road loss t… | How many yards was the second longest… | 42 | FAIL | DROP |
| 52db14ed-5426-46ec-b0ae-4ef843b2d692 | Hoping to rebound from their tough… | How long was the last touchdown? | 18-yard | FAIL | DROP |
| 31b36417-aad1-412c-b0e5-9c1faaed233f | As of the census of 2000, there were… | How many in percent from the census… | 87.1 | FAIL | DROP |
| 57d6f476-5163-4069-baab-b53504a3e662 | In the United States, conscription began i… | How many more men were registered in… | 15000000 | FAIL | DROP |

‹ Previous 1 2 3 ... 149 Next ›

https://arxiv.org/pdf/2407.16833

# Optional) Long Context vs RAG?


Figure 1: While long-context LLMs (LC) surpass RAG in long-context understanding, RAG is significantly more cost-efficient. Our approach, SELF-ROUTE, combining RAG and LC, achieves comparable performance to LC at a much lower cost.

https://www.databricks.com/blog/long-context-rag-performance-llms
https://www.databricks.com/blog/long-context-rag-capabilities-openai-o1-and-google-gemini
https://huggingface.co/datasets/PatronusAI/financebench/

# Optional) RAG 에서 Generation 이 문제인 경우

| # Retrieved chunks | 1 | 5 | 13 | 29 | 61 | 125 | 189 | 253 | 317 | 381 |
|---|---|---|---|---|---|---|---|---|---|---|
| Recall@k \ Context Length | 2k | 4k | 8k | 16k | 32k | 64k | 96k | 128k | 160k | 192k |
| Databricks DocsQA | 0.547 | 0.856 | 0.906 | 0.957 | 0.978 | 0.986 | **0.993** | **0.993** | **0.993** | **0.993** |
| FinanceBench | 0.097 | 0.287 | 0.493 | 0.603 | 0.764 | 0.856 | **0.916** | **0.916** | **0.916** | **0.916** |
| NQ | 0.845 | 0.992 | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| HotPotQA | 0.382 | 0.672 | 0.751 | 0.797 | 0.833 | 0.864 | 0.880 | **0.890** | **0.890** | **0.890** |
| Average | 0.468 | 0.702 | 0.788 | 0.839 | 0.894 | 0.927 | 0.947 | **0.95** | **0.95** | **0.95** |

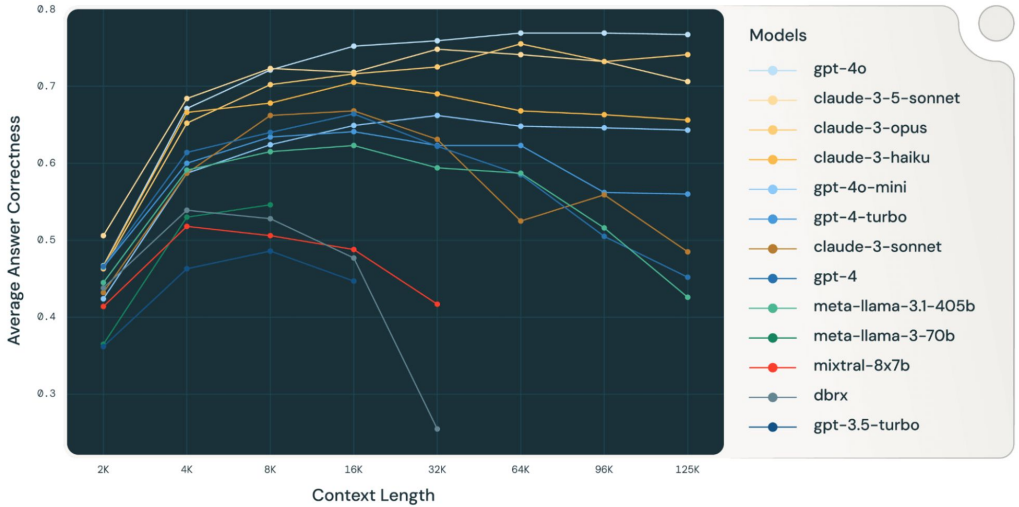



Figure 1: Long context performance of GPT, Claude, Llama, Mistral and DBRX models on 4 curated RAG datasets (Databricks DocsQA, FinanceBench, HotPotQA and Natural Questions)

# Optional) Multi hop question (HotPotQA style)

**Question**: *"Which company was founded by the author of the book 'The Road Ahead'?"*

### Steps to answer:

1. **First hop**: Identify the author of the book "The Road Ahead."
    - The author is *Bill Gates*.
2. **Second hop**: Find out which company Bill Gates founded.
    - Bill Gates is the co-founder of *Microsoft*.

**Answer**: *Microsoft*.