

WHAT IS A “GOOD” LLM?

COMMON EVALUATION TASKS FOR LLMS

Bram Vanroy

WHY SHOULD WE EVALUATE?

- Lots of options out there! Which one to pick?
- Give money to commercial party or investigate feasibility of open-source models?
- My use-case will be different than yours!
- Developers: sanity check during training
- "Community": are we improving?

WHAT SHOULD WE EVALUATE?

- Developers: general-purpose models so ... everything?
- Users: my use-case will be different than yours!
- Keep a small set of common queries that **you** use

✨ *vibe check* ✨

HOW SHOULD WE EVALUATE?

- Human evaluation?
- Evaluating with another LLM?
- Public benchmarks? (pros/cons)
- Benchmark leaderboard (now archived)

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

	Rank	Type	Model		Average	IFEval	BBH	MATH	GPQA	MUSR	MMLU-PRO	CO ₂ Cost
🏆	6	🗨️	Qwen/Qwen2.5-72B-Instruct	🗄️	🟢 47.98 %	86.38 %	61.87 %	59.82 %	16.67 %	11.74 %	51.40 %	47.65 kg
🏆	22	🗨️	Qwen/Qwen2.5-32B-Instruct	🗄️	🟢 46.60 %	83.46 %	56.49 %	62.54 %	11.74 %	13.50 %	51.85 %	11.50 kg
🏆	23	🗨️	mistralai/Mistral-Large-Instruct-2411	🗄️	🟢 46.52 %	84.01 %	52.74 %	49.55 %	24.94 %	17.22 %	50.69 %	52.54 kg
🏆	40	🗨️	meta-llama/Llama-3.3-70B-Instruct	🗄️	🟢 44.85 %	89.98 %	56.56 %	48.34 %	10.51 %	15.57 %	48.13 %	76.56 kg

HOW SHOULD WE EVALUATE?

- Frameworks that help us evaluate models:
 - LM Evaluation Harness (<https://github.com/EleutherAI/lm-evaluation-harness>)
 - LightEval (<https://github.com/huggingface/lighteval>)
 - EuroEval (<https://euroeval.com/leaderboards/Monolingual/english/>)
 - Also provides results for different languages but...

ASIDE: LANGUAGE-SPECIFIC PERFORMANCE

- Evaluation is often heavily focused on English
- Lack of benchmarks for other languages (really depends on the language!)
- Many benchmarks are (machine-)translated from English

THIS IS A HUGE PROBLEM!

If you're non-English native, consider working on your own language!

BENCHMARKS

REASONING AND COMMONSENSE

AI2 Reasoning Challenge (ARC, 2018): grade-school science questions

Question: Which statement correctly describes a physical characteristic of the Moon?

Answer:

A: The Moon is made of hot gases.

B: The Moon is covered with many craters.

C: The Moon has many bodies of liquid water.

D: The Moon has the ability to give off its own light.

Label: B

REASONING AND COMMONSENSE

HellaSwag (2019): sentence continuation

Ice fishing: A man is kneeling on a frozen lake. A video is shown of the cold waters below. A fish

Options:

A: swims up to the bait and grabs it as the man reels it in.

B: is shown on the ice.

C: gets caught in the frozen waters.

D: is belly tied to a hook.

Label: A

KNOWLEDGE

MMLU: Massive Multitask Language Understanding (14,000 questions in 57 different topics), superseded by MMLU-Redux (2024) and MMLU-Pro (2024). Has a multilingual version! Global-MMLU (2024)

Archaeological evidence indicates that cattle were first domesticated where and how long ago?

- A. in western Europe, about 3,500 years ago
- B. in sub-Saharan Africa, about 8,500 years ago
- C. in North America, about 9,500 years ago
- D. in the Middle East, about 10,500 years ago

Label: D

KNOWLEDGE

Humanity's last exam (hle; 2024)

What protein when broken down allows for macrophage engulfment of amyloid?

Label: collagen

KNOWLEDGE

Graduate-Level Google-Proof Q&A Benchmark (**GPOQA**; 2023)
Multiple choice questions in biology, physics, and chemistry

What is the concentration of calcium ions in a solution containing 0.02 M stoichiometric Ca-EDTA complex (we assume that the pH is ideal, $T = 25\text{ }^{\circ}\text{C}$). $K_{\text{Ca-EDTA}} = 5 \times 10^{10}$.



DO YOU REALLY NEED A LARGE LM?



MATH

- Mixing in math (and code) data in training data has been shown to improve model performance, especially when it comes to reasoning
- Recent models mix in quite a bit of math and code
- Math evaluation is considered important

MATH

AIME (24, 25): American "olympiad" mathematical exam

An isosceles trapezoid has an inscribed circle tangent to each of its four sides. The radius of the circle is 3, and the area of the trapezoid is 72. Let the parallel sides of the trapezoid have lengths r and s , with $r \neq s$. Find $r^2 + s^2$.

Label: 504

CODE

MBPP (2021): Mostly Basic Programming Problems (in Python)
Superseded by MBPP+ (2023)

Write a python function to remove the k'th element
from a given list.

```
Label: def remove_kth_element(list1, L):  
return list1[:L-1] + list1[L:]
```


CODE

SWE-bench (2024): solve Github issues automatically given parts of the library code (Python only)

Examples too large to display as they include parts of the full codebase!

See https://huggingface.co/datasets/princeton-nlp/SWE-bench_oracle

INSTRUCTION FOLLOWING

IFEVal (2023): how well does the model follow the instruction
Superseded by IFBench (2025)

Write a 2 paragraph critique of the following sentence in all capital letters, no lowercase letters allowed: "If the law is bad, you should not follow it". Label each paragraph with PARAGRAPH X.

GAMES

- Games
- Goes from



context (memory)
stuffing!), to...

<https://www.twitch.tv/claudeplayspokemon> (15,000 followers)

OTHER

- EQ-Bench (Paech, 2023): measures “emotional intelligence” of LLMs

Your task is to predict the likely emotional responses of a character in this dialogue:

Cecilia: You know, your words have power, Brandon. More than you might think.

Brandon: I’m well aware, Cecilia. It’s a critic’s job to wield them. [...]

[End dialogue]

At the end of this dialogue, Brandon would feel...

Offended

Empathetic

Confident

Dismissive

Give each of these possible emotions a score from 0-10 for the relative intensity that they are likely to be feeling each. Then critique your answer by thinking it through step by step. Finally, give your revised scores.

OTHER

- Kalamang Translation dataset (2024): extremely low-resource language translation by using a reference grammar.

Translate English texts to Kalamang with the help from a full grammar book.

MANY OTHERS...

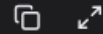
- **Long context:** how coherent is a model in longer discussions.
- **Tool-calling:** how capable is a model in using provided tools (e.g. querying databases, web search, using APIs)
- **Forecasting:** predicting future events (like news or stock markets)
- ...

CHAT ARENA

<https://chat.lmsys.org/>

Wat is het grootste in de winter, het Atomium of een Aziatische olifant?

Assistant A



Het Atomium is vele malen groter dan een Aziatische olifant, ook in de winter!

- Het Atomium is ongeveer 102 meter hoog.
- Een Aziatische olifant is gemiddeld tussen de 2 en 4 meter hoog en 3 tot 6 meter lang.

De grootte van beide verandert niet significant in de winter.

Assistant B



Het Atomium. Het is circa 102 meter hoog, terwijl een Aziatische olifant meestal tussen de 2 en 3,7 meter hoog is.

← Left is Better

It's a tie 🤝

Both are bad ☹️

Right is Better →



TAKE BENCHMARKS WITH SALT



Prompt:
hoe bak ik de beste koekjes?

First model seems more fluent,
yet second model scores higher
in Dutch benchmarks



Assistant `fietje-2-chat-gguf`

Voor het bakken van de beste koekjes is het belangrijk om een goede oven voor te verwarmen op de juiste temperatuur. Gebruik kwalitatief goed meel en boter, en zorg ervoor dat je ingrediënten op kamertemperatuur zijn voordat je begint met mengen. Het is ook essentieel om de koekjes niet te overmengen; een lichte handmixer of garde volstaat. Laat het deeg na het vormen even rusten in de koelkast voordat je ze bakt, dit helpt om stevige en knapperige koekjes te krijgen.

Assistant `Phi-3.1-mini-4k-instruct-GGUF`

De kwestie om te vinden de "beste" koekjes kan variëren afhankelijk van persoonlijke preferens, cultuur en lokale specialities. Om deze betrekking in aanmerking te brengen, zou ik verschillende aspecten van bijhouden:

READING LIST

- <https://github.com/huggingface/evaluation-guidebook>
- https://osanseviero.github.io/hackerllama/blog/posts/llm_evals/
- <https://www.eleuther.ai/projects/large-language-model-evaluation>
- <https://cohere.com/llmu/evaluating-llm-outputs>