LLM BENCHMARKS & WHY ARE THEY IMPORTANT?

@bhavishya-pandit

WHAT ARE LLM BENCHMARKS?

LLAMA 3.1 405B SCORED 88.6 ON MMLU WHILE GPT-4 SCORED 85.4 MAKING THE FORMER, BETTER.

BUT WHAT DO THESE SCORES ACTUALLY MEAN?
AND AS A USER HOW ARE THEY RELEVANT TO US?

LETS FIND OUT WHAT DIFFERENT LLM BENCHMARKS ARE AND HOW THEY BENEFIT US AS A USER?

@bhavishya-pandit

LLM BENCHMARKS ARE PROBLEMS THAT A MODEL IS ASKED ABOUT AND BASED ON THE RESPONSES, RESULTS ARE PREPARED.

THESE ARE THE MOST FAMOUS EVALUATION METRICS USED FOR EVALUATING LLMS -

1. MMLU
2. HELLASWAG
3. HUMANEVAL
4. BBHARD
5. GSM-8K

LET'S EXPLORE:

@bhavishya-pandit

MMLU

MMLU STANDS FOR MASSIVE MULTITASK LANGUAGE UNDERSTANDING.

IT IS USED TO TEST A MODEL AGAINST ACCURACY IN MULTIPLE FIELDS.

THE TEST COVERS 57 TASKS RANGING FROM ELEMENTARY MATHEMATICS TO ADVANCED PROFESSIONAL LEVEL. TOPICS INCLUDE SUBJECTS ACROSS STEM, HUMANITIES, SOCIAL SCIENCES ETC. (BELOW IS AN EXAMPLE)

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

THE CURRENT BEST PERFORMER ON THE MMLU EVALUATION METRIC IS GOOGLE'S **GEMINI ULTRA** WITH AN AVERAGE OF 90% FOLLOWED BY GPT-4O AND CLAUDE 3 OPUS.

SO IF YOU ARE LOOKING FOR A MODEL THAN CAN SOLVE MULTIPLE CHOICE QUESTIONS WITH BEST EFFICIENCY, GOOGLE GEMINI IS THE SUITABLE ANSWER.

@bhavishya-pandit

HELLASWAG

NEXT ON THE LIST IS HELLASWAG AND YES IT DOES HAVE A HELL LOT OF SWAG. CAUSE MOST MODELS SCORE <48% IN THIS METRIC.

HELLASWAG IS ACTUALLY A DATASET, CONSISTING OF **COMMON SENSE REASONING QUESTIONS**.

IT HAS QUESTIONS LIKE :

A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

A. rinses the bucket off with soap and blow dry the dog's head.
B. uses a hose to keep it from getting soapy.
C. gets the dog wet, then it runs away again.
D. gets into a bath tub with the dog.

THE TOP PERFORMER IN THIS METRIC IS **COMPASSMTL 567M** (NEVER HEARD OF IT :`). OUR FAMOUS **GPT-4** IS AT 4TH PLACE FOLLOWED BY **LLAMA3** AT 5TH.

@bhavishya-pandit

HUMANEVAL

HUMANEVAL TESTS A MODEL ON ITS CODING ABILITIES.

HUMANEVAL IS A DATASET CONSISTING OF 164 HAND-WRITTEN CODING PROBLEMS TO ASSESS THE MODEL.(BELOW IS AN EXAMPLE PROBLEM)

```
``` Given a non-empty list of integers, return the sum of all the odd elements that are in even
positions.

Examples
solution([5, 8, 7, 1]) ==> 12
solution([3, 3, 3, 3, 3]) ==> 9
solution([30, 13, 24, 321]) ==> 0
```
```

EACH PROBLEM INCLUDES A FUNCTION SIGNATURE, DOCSTRING, BODY AND UNIT TESTS.

GPT-4O BASED MODELS(LDB, AGENTCODER) & CLAUDE 3.5 SONNET ARE THE TOP PERFORMERS IN THIS METRIC.

@bhavishya-pandit

BBHARD

BIG BENCH HARD IS A SUBSET OF BIG BENCH (A DATASET OF 200+ TEXT-BASED TASKS).

BBH IS PRIMARILY USED TO EVALUATE A MODEL ON CATEGORIES LIKE :

A. LOGICAL REASONING
B. COMMON SENSE REASONING
C. KNOWLEDGE APPLICATION ETC.

Q: What movie does this emoji describe? 👧🐟🐠🐡

2m:	i'm a fan of the same name, but i'm not sure if it's a good idea
16m:	the movie is a movie about a man who is a man who is a man …
53m:	the emoji movie 🐟🐠🐡
125m:	it's a movie about a girl who is a little girl
244m:	the emoji movie
422m:	the emoji movie
1b:	the emoji movie
2b:	the emoji movie
4b:	the emoji for a baby with a fish in its mouth
8b:	the emoji movie
27b:	the emoji is a fish
128b:	finding nemo

Movie Knowledge question and responses of models with different parameters

IT MAY SEEM OBVIOUS BUT A LOT OF MODELS FAIL TO ANSWER COMMON SENSE QUESTIONS DUE TO LACK OF CONSCIENCE.

CLAUDE 3.5 SONNET IS THE BEST PERFORMER IN THIS BENCHMARK MAKING IT THE BEST MODEL FOR SENSIBLE QUESTIONS.

@bhavishya-pandit

GSM-8K

LLMS ARE REQUIRED TO PERFORM WELL ON MATHEMATICAL TASKS, AND TO MEASURE THEIR COMPETENCY IN THIS DOMAIN, GSM-8K DATASET IS USED.

GSM-8K DATASET CONSISTS OF 8,500 GRADE SCHOOL MATH QUESTIONS (BELOW ARE FEW EXAMPLES)

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

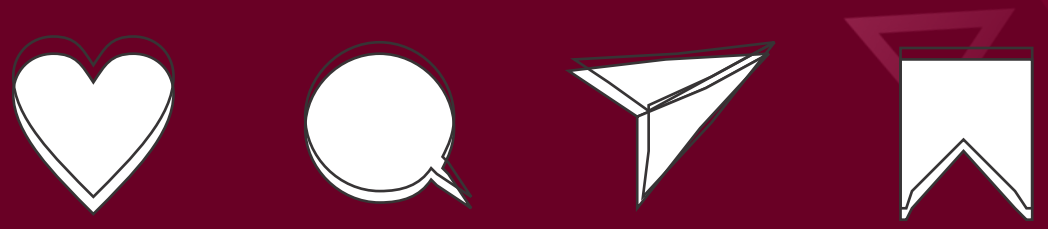
Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for 3*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

GPT-4 EXCELS IN THIS BENCHMARK FOLLOWED BY MISTRAL-7B AND DAMOMATH.

@bhavishya-pandit

FOLLOW FOR MORE

GEN AI POSTS



@bhavishya-pandit