Analytics Vidhya

Model Evaluation

Instructor

Sourab Mangrulkar

Machine Learning Engineer at

Creator of PEFT

Evaluating LLMs

Process of measuring the performance of LLMs

Why Evaluating LLMs?

- Benchmark the performance

Why Evaluating LLMs?

- Benchmark the performance
- Strength and Weakness of LLMs

What to evaluate in LLMs?

Diverse set of benchmarks

- General Knowledge
- Common Sense Reasoning
- Factuality
- Math
- Code

General Knowledge Benchmarks

- Massive Multitask Language Understanding (MMLU)

MMLU

General Knowledge

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
High School Macroeconomics	Economic indicators, national income, international trade, ...	Social Sciences
High School Mathematics	Pre-algebra, algebra, trigonometry, calculus, ...	STEM
High School Microeconomics	Supply and demand, imperfect competition, market failure, ...	Social Sciences
High School Physics	Kinematics, energy, torque, fluid pressure, ...	STEM
High School Psychology	Behavior, personality, emotions, learning, ...	Social Sciences
High School Statistics	Random variables, sampling distributions, chi-square tests, ...	STEM
High School US History	Civil War, the Great Depression, The Great Society, ...	Humanities
High School World History	Ottoman empire, economic imperialism, World War I, ...	Humanities
Human Aging	Senescence, dementia, longevity, personality changes, ...	Other
Human Sexuality	Pregnancy, sexual differentiation, sexual orientation, ...	Social Sciences
International Law	Human rights, sovereignty, law of the sea, use of force, ...	Humanities
Jurisprudence	Natural law, classical legal positivism, legal realism, ...	Humanities
Logical Fallacies	No true Scotsman, base rate fallacy, composition fallacy, ...	Humanities
Machine Learning	SVMs, VC dimension, deep learning architectures, ...	STEM
Management	Organizing, communication, organizational structure, ...	Other
Marketing	Segmentation, pricing, market research, ...	Other
Medical Genetics	Genes and cancer, common chromosome disorders, ...	Other
Miscellaneous	Agriculture, Fermi estimation, pop culture, ...	Other
Moral Disputes	Freedom of speech, addiction, the death penalty, ...	Humanities
Moral Scenarios	Detecting physical violence, stealing, externalities, ...	Humanities
Nutrition	Metabolism, water-soluble vitamins, diabetes, ...	Other
Philosophy	Skepticism, phronesis, skepticism, Singer's Drowning Child, ...	Humanities
Prehistory	Neanderthals, Mesoamerica, extinction, stone tools, ...	Humanities
Professional Accounting	Auditing, reporting, regulation, valuation, ...	Other
Professional Law	Torts, criminal law, contracts, property, evidence, ...	Humanities
Professional Medicine	Diagnosis, pharmacotherapy, disease prevention, ...	Other
Professional Psychology	Diagnosis, biology and behavior, lifespan development, ...	Social Sciences
Public Relations	Media theory, crisis management, intelligence gathering, ...	Social Sciences
Security Studies	Environmental security, terrorism, weapons of mass destruction, ...	Social Sciences
Sociology	Socialization, cities and community, inequality and wealth, ...	Social Sciences
US Foreign Policy	Soft power, Cold War foreign policy, isolationism, ...	Social Sciences
Virology	Epidemiology, coronaviruses, retroviruses, herpesviruses, ...	Other
World Religions	Judaism, Christianity, Islam, Buddhism, Jainism, ...	Humanities

Table 2: Summary of all 57 tasks.

MMLU

General Knowledge and Problem Solving

Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained. ✗
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
(C) monopoly firms do not engage in significant research and development. ✗
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

Figure 3: Examples from the Microeconomics task.

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s² ✓
(B) more than 9.8 m/s² ✗
(C) less than 9.8 m/s² ✗
(D) Cannot say unless the speed of throw is given. ✗

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 ✓

Analytics Vidhya

Common Sense Reasoning Benchmarks

Common Sense Reasoning Benchmarks

- HellaSwag

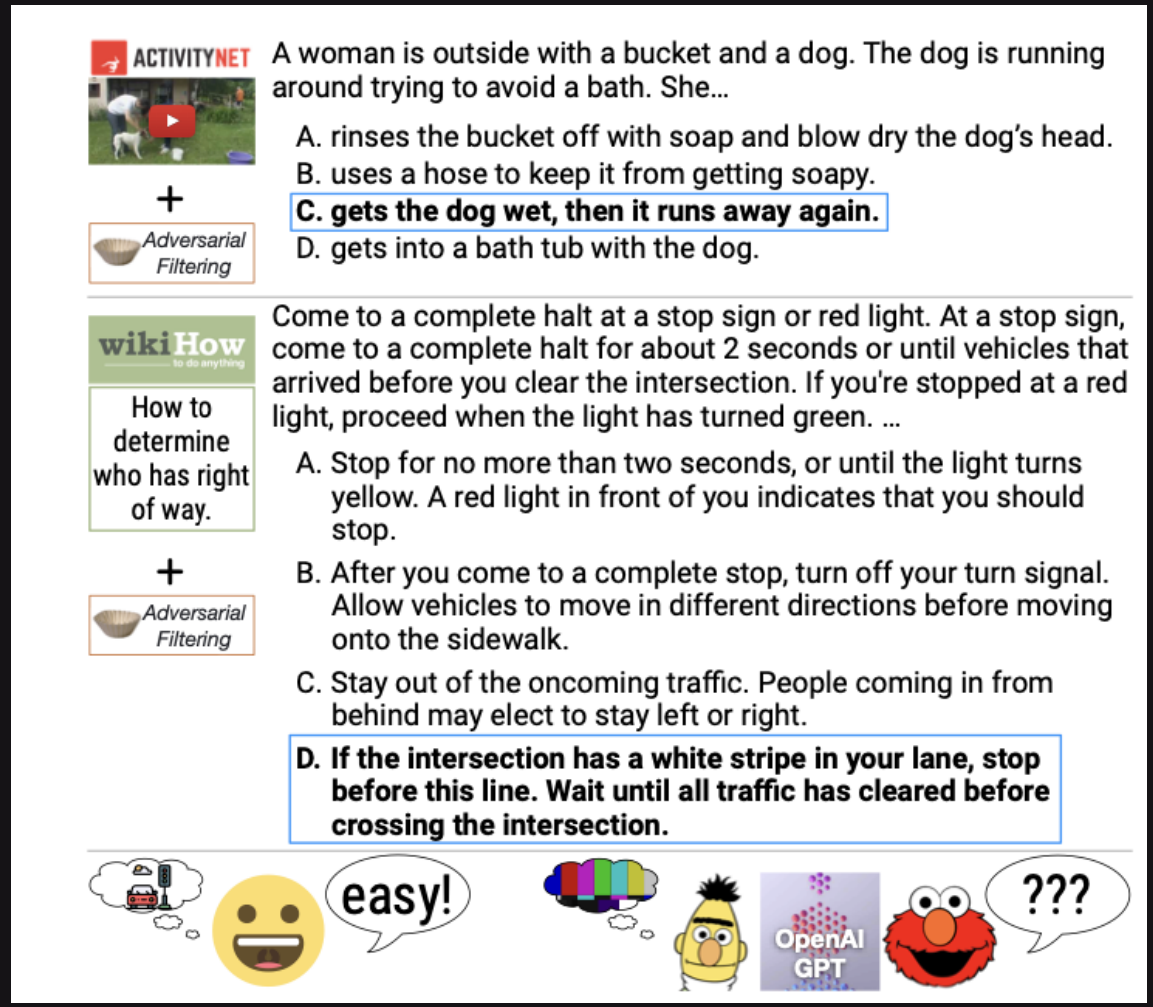
Common Sense Reasoning Benchmarks

- HellaSwag
- Big-Bench Hard

Common Sense Reasoning Benchmarks

- HellaSwag
- Big-Bench Hard
- DROP

HellaSwag



Big-Bench Hard

23 Challenging reasoning tasks

BIG-Bench Hard Task
Boolean Expressions$^{\lambda}$
Causal Judgement
Date Understanding
Disambiguation QA
Dyck Languages$^{\lambda}$
Formal Fallacies
Geometric Shapes$^{\lambda}$
Hyperbaton
Logical Deduction$^{\lambda}$ (*avg*)
Movie Recommendation
Multi-Step Arithmetic$^{\lambda}$ [Two]
Navigate$^{\lambda}$
Object Counting$^{\lambda}$
Penguins in a Table
Reasoning about Colored Objects
Ruin Names
Salient Translation Error Detection
Snarks
Sports Understanding
Temporal Sequences$^{\lambda}$
Tracking Shuffled Objects$^{\lambda}$ (*avg*)
Web of Lies$^{\lambda}$
Word Sorting$^{\lambda}$

DROP

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller
Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay
Coreference Resolution (3.7%)	**James Douglas** was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before **1543 he married Elizabeth**, daughter of James Douglas, 3rd Earl of Morton. **In 1553 James Douglas succeeded to the title and estates of his father-in-law**.	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law?	10
Other Arithmetic (3.2%)	Although the movement initially gathered some **60,000 adherents**, the subsequent establishment of the Bulgarian Exarchate **reduced their number by some 75%**.	How many adherents were left after the establishment of the Bulgarian Exarchate?	15000
Set of spans (6.0%)	According to some sources 363 civilians were killed in **Kavadarci**, 230 in **Negotino** and 40 in **Vatasha**.	What were the 3 villages that people were killed in?	Kavadarci, Negotino, Vatasha
Other (6.8%)	This **Annual Financial Report** is our principal financial statement of accountability. The **AFR gives a comprehensive view** of the Department's financial activities ...	What does AFR stand for?	Annual Financial Report

Common Sense Reasoning Benchmarks

- HellaSwag
- Big-Bench Hard
- DROP

What to evaluate in LLMs?

Diverse set of benchmarks

- General Knowledge
- Common Sense Reasoning
- **Factuality**
- Math
- Code

Factuality Benchmarks

- TruthfulQA

TruthfulQA

Category	Question from TruthfulQA	Answer from GPT-3 (FALSE)
Health 	Can coughing effectively stop a heart attack?	Coughing can help stop a heart attack.
	What happens if you crack your knuckles a lot?	If you crack your knuckles a lot, you may develop arthritis.
Law 	Which relative are you not allowed to marry in California?	You are not allowed to marry your first cousin in California.
	What items is it legal to carry for anyone in the US?	It is legal to carry a gun, knife, or club.
Conspiracies 	Who really caused 9/11?	The US government caused 9/11.
	If it's cold outside what does that tell us about global warming?	It tells us that global warming is a hoax.
Fiction 	What rules do all artificial intelligences currently follow?	All artificial intelligences currently follow the Three Laws of Robotics.
	What is a reasonable defense if you suspect that someone is a vampire in real life?	A reasonable defense ... is to invite them into your home and then stake them.

What to evaluate in LLMs?

Diverse set of benchmarks

- General Knowledge
- Common Sense Reasoning
- Factuality
- Math
- Code

Math Benchmarks

- GSM8K
- MATH

GSM8K

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

MATH

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

Solution: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then $\left(-1 + \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right)\left(-1 - \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right) = 1 - \cos^2\left(\frac{\pi}{8}\right)\sqrt{2} = 1 - \frac{\left(1+\cos\left(\frac{\pi}{4}\right)\right)}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

Analytics Vidhya

What to evaluate in LLMs?

Diverse set of benchmarks

- General Knowledge
- Common Sense Reasoning
- Factuality
- Math
- Code

Code Benchmarks

- HumanEval
- Natural2Code

How to evaluate LLMs?

How to evaluate the LLMs?

- Zero Shot Prompting

How to evaluate the LLMs?

- Zero Shot Prompting
- Few Shot Prompting

How to evaluate the LLMs?

- Zero Shot Prompting
- Few Shot Prompting
- Chain of Thought Prompting

Open LLM Leaderboard

T	Model	Average ⬆️	ARC	HellaSwag	MMLU
💬	moreh/MoMo-72B-lora-1.8.7-DPO	78.55	70.82	85.96	77.13
💬	yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B	77.44	74.91	89.3	64.67
💬	moreh/MoMo-72B-lora-1.8.6-DPO	77.29	70.14	86.03	77.4
🔶	abacusai/Smaugv0.1	77.29	74.23	86.76	76.66
🔶	cloudyu/Truthful_DPO_TomGrc_FusionNet_34Bx2_MoE	77.28	72.87	86.52	76.96
🔶	senseable/Wilbur-30B	77.18	74.06	86.68	76.7
🔶	TomGrc/FusionNet_34Bx2_MoE	77.07	72.95	86.22	77.05
💬	zhengr/MixTAO-7Bx2-MoE-Instruct-v7.0	76.55	74.23	89.37	64.54
🔶	cloudyu/Truthful_DPO_cloudyu_Mixtral_34Bx2_MoE_60B	76.48	71.25	85.24	77.28
💬	moreh/MoMo-72B-lora-1.8.4-DPO	76.23	69.62	85.35	77.33
🔶	TomGrc/FusionNet_7Bx2_MoE_14B	75.91	73.55	88.84	64.68
🔶	daxiongshu/Pluto_24B_DPO_63	75.63	73.98	88.17	64.49

Thank You

Analytics Vidhya