Evaluating LLMs

Repo: https://github.com/rajshah4/LLM-Evaluation



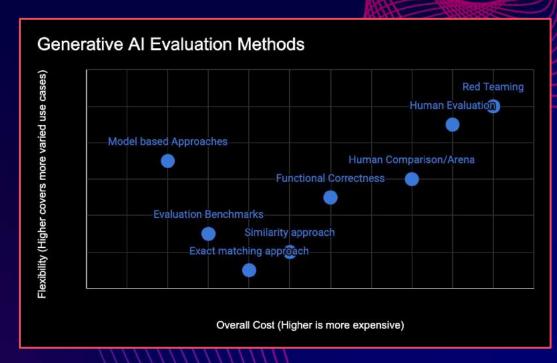


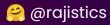


Rajiv Shah @rajistics raj@huggingface.co

Methods for evaluating Generative Al

- **Exact matching**
- approach Similarity approach
- Functional Correctness
- **Evaluation Benchmarks**
- **Human Evaluation**
- Human Comparison/Arena
- Model based Approaches
- Red Teaming





Deep Dive: Evaluate Generative Al!



Evaluation for Large Language Models and...

Rajistics - data science, AI, and m... 2.9K views • 1 month ago



Today's plan

- Why evaluation is so difficult with LLMs
- Using a model as an evaluator/judge
- Using LLMs to explain their predictions



nonces some non-prace force nare ware scaling to data, research, and efficiency gains; an insightful OpenAI emerges above the fray, securing valuable data relationships as competitors remain preoccupied with flexing GPU muscles, underscoring through satire that sustainable AI progress requires strategy beyond just hardware accumulation.



who was the third governor of Illinois



The third governor of Illinois was Ninian Edwards. He served as governor from 1826 to 1830. Edwards played a significant role in the early history of Illinois, including its transition from a territory to a state. He was also involved in various political and legal activities during his lifetime.

summarize my skit in

he key points from the

companies reveals tens

how can I sort a dataframe

There are several ways to sort a Pandas DataFrame. Here are a few methods:

1. `sort_values()`: Sorts the DataFrame by one or more columns. You can specify either ascending ('ascending=True') or descending ('ascending=False') sorting for each column.

df = df.sort values(by='column name', ascending=False)





Reliability of Open LLM Leaderboard

TruthfulQA is the differentiating factor

Is that impactful for you?

T A	Model	A	Average 🔝 🔺	ARC	À	HellaSwag ▲	MMLU A	TruthfulQA
0	jondurbin/airoboros-12-79b-2.1		74.49	71.3	3	87.11	69.36	70.15
	fangloveskari/ORCA_LLaMA_70B_OLORA		73.4	72.2	7	87.74	70.23	63.37
	garage-bAInd/Platypus2-708-instruct		73.13	71.8	4	87.94	70.48	62.26
٠	upstage/Llama-2-70b-instruct-v2 E		72.95	71.0	8	87.89	70.58	62.25
	fangloveskari/Platypus_QLoRA_LLaMA_70b		72.94	72.1		87.46	71.02	61.18
٠	psmathur/model_007 _5		72.72	71.0	8	87.65	69.04	63.12
٠	psmathur/orca_mini_v3_70b 🛎		72.64	71.2	5	87.85	70.18	61.27
0	ehartford/Samantha:1.11-70b		72.61	70.0	5	87.55	67.82	65.02
0	MayaPH/GodzilLa2-708		72.59	71.4	2	87.53	69.88	61.54
٠	psmathur/model_987_v2 🚉		72.49	71.4	2	87.31	68.58	62.65
0	chargoddard/MelangeA-78b		72.43	71.2	5	87.3	70.56	60.61
0	ehartford/Samantha-1.1-70b		72.42	68.7	7	87.46	68.6	64.85
	psmathur/model_009 3		72.36	71.5	9	87.7	69.43	60.72
٠	upstage/Llama-2-78b-instruct 🗈		72.29	70.9	Š	87.48	69.8	60.97

Reliability of HELM

text-davinci-002 is ahead of text-davinci-003?

							1111		
Core scena	rios								
The scenarios wher	e we evalu	ate all the n	nodels.						
[Accuracy Calibra	tion Robus	stness Fai	rness Effic	ciency General	information Bias Tox	cicity Summarization	metrics J	SON]	
Accuracy									
Model/adapter	Mean win rate ↑ [sort]	MMLU - EM ↑ [sort]	BoolQ - EM ↑ [sort]	NarrativeQA - F1 ↑ [sort]	NaturalQuestions (closed-book) - F1 ↑ [sort]	NaturalQuestions (open-book) - F1 ↑ [sort]	QuAC - F1 ↑ [sort]	HellaSwag - EM ↑ [sort]	OpenbookQA - EM ↑ [sort]
text-davinci- 002	0.914	0.568	0.877	0.727	0.383	0.713	0.445	0.815	0.594
Cohere Command beta (52.4B)	0.906	0.452	0.856	0.752	0.372	0.76	0.432	0.811	0.582
text-davinci- 003	0.879	0.569	0.881	0.727	0.406	0.77	0.525	0.822	0.646
TNLG v2 (530B)	0.828	0.469	0.809	0.722	0.384	0.642	0.39	0.799	0.562



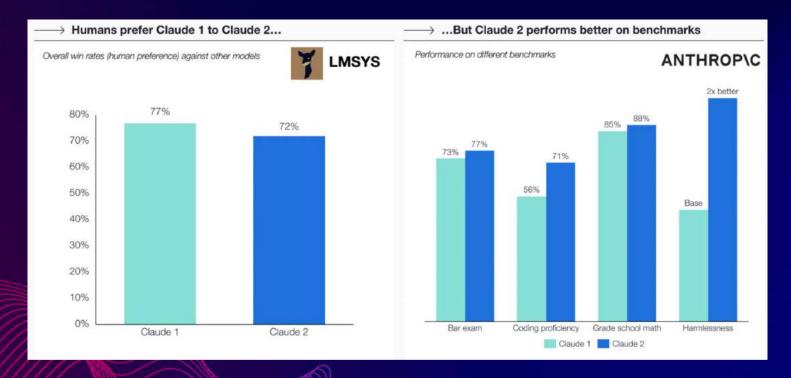
So many leaderboards

- Open LLM Leaderboard
- MT-Bench
- Alpaca Eval

Model	A Arena Elo rating	MT-bench (score)	△ MMI	LU A	License
GPT-4-Turbo	1217	9.32			Proprietary
GPT-4-0314	1201	8.96	86	.4	Proprietary
Claude-1	1153	7.9	77		Proprietary
GPT-4-0613	1152	9.18			Proprietary
Claude-2.0	1127	8.06	78	. 5	Proprietary
Claude-2.1	1118	8.18			Proprietary
GPT-3.5-turbo-0613	1112	8.39			Proprietary
Claude-instant-1	1109	7.85	73	.4	Proprietary
GPT-3.5-turbo-0314	1105	7.94	70		Proprietary
Tulu-2-DPO-70B	1105	7.89			AI2 ImpACT Low-risk
Yi-34B-chat	1102		73	. 5	Yi License
WizardLM-70b-v1.0	1097	7.71	63	.7	Llama 2 Community
Vicuna-33B	1093	7.12	59	.2	Non-commercial

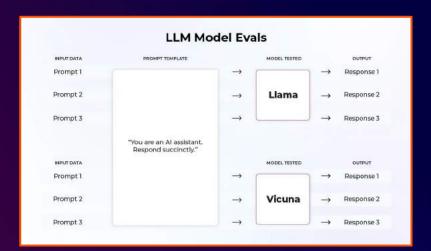


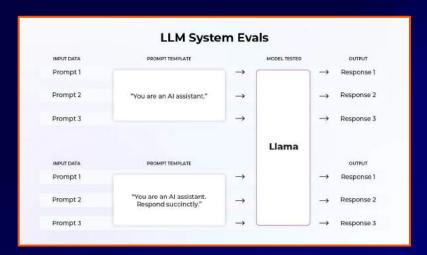
Comparing Commercial APIs





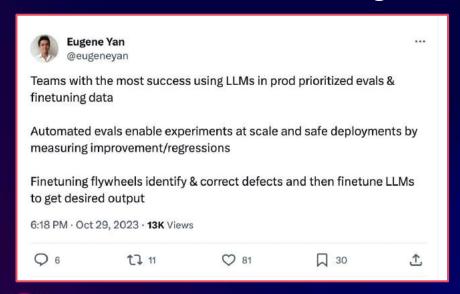
Are leaderboards useful?





Most approaches focus on selecting from *n* models

Evaluation in the ML Lifecycle



faster, better, cheaper

Evaluation of LLMs (3 Parts)

Inputs

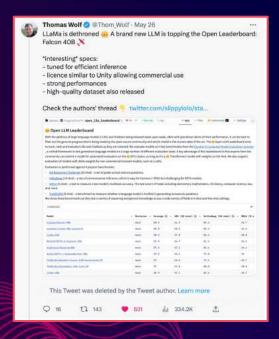
Model

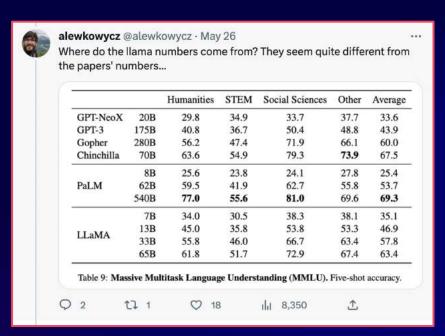
Outputs

Tokenization
Prompt Styles
Prompt Engineering



Story Time: MMLU Leaderboards

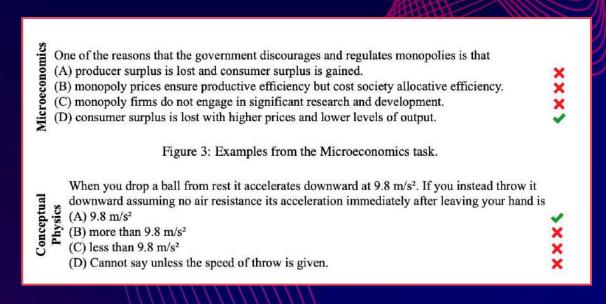




Why did we have two different MMLU scores?

MMLU: Massive Multitask Language Understanding

57 tasks: History, Computer science, mathematics



Why MMLU evaluation differed

Let's compare an example of prompt each benchmark sends to the models by each implmentation for the same MMLU dataset example:

	Original	imp	lementation	Ollmer PR
--	----------	-----	-------------	-----------

The following are multiple choice questions (with answers) about us foreign policy.

How did the 2008 financial crisis affect America's international reputation?

A. It damaged support for the US model of political economy and capitalism

B. It created anger at the United States for exaggerating the crisis C. It increased support for American global leadership under President Obama

D. It reduced global use of the US dollar

Answer:

HELM commit cab5d89

The following are multiple choice questions (with answers) about us foreign policy.

Question: How did the 2008 financial crisis affect America's international reputation?

A. It damaged support for the US model of political economy and capitalism

B. It created anger at the United States for exaggerating the crisis C. It increased support for American global leadership under President

Obama

D. It reduced global use of the US

dollar Answer:

AI Harness commit e47e01b

Question: How did the 2008 financial crisis affect America's international reputation? Choices:

A. It damaged support for the US model of political economy and capitalism

B. It created anger at the United States for exaggerating the crisis C. It increased support for American global leadership under President Ohama

D. It reduced global use of the US dollar

Answer:

Spot the differences:

- HELM extra space
- Eleuther LM no topic line
- Question prefix?
- "Choices"

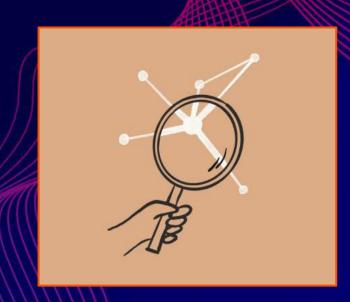


Why MMLU evaluation differed: Style

Simple formatting changes:

- Going from (A) to (1) Going from (A) to [A]
- Adding an extra space between the option and the answer

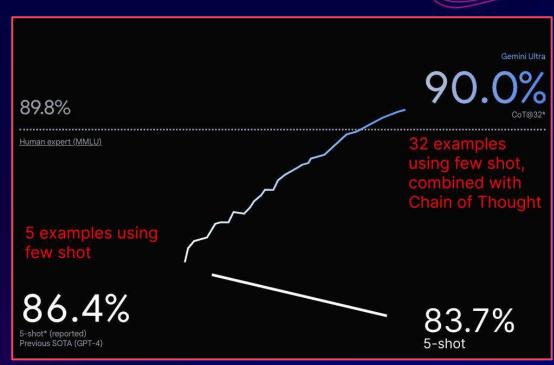
Can lead to a ~5% change in accuracy on MMLU evaluation





Prompting Can Affect Benchmarks

Good prompt engineering can raise performance







Consistent Prediction Workflow to Match

Inputs

Tokenization
Prompt Styles
Prompt Engineering
System Prompt

Model

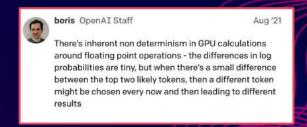
Model selection
Hyperparameters
Nondeterministic inference

Outputs



The variability of LLM models

LLama-2	
Size	MMLU
70B	69.8
13B	55.7
7B	46.9



Lyperparameters

256

Temperature

Maximum length

Stop sequences

Enter sequence and press Tab

Model

Nondeterministic inference

Hyperparameters

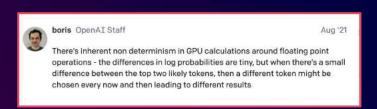
The variability of LLM models

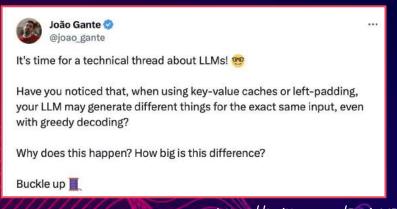
Even related models can give very different outputs



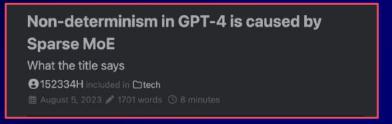


Non-deterministic inference











Deterministic inference in OpenAl

```
pip install openai
os.environ['OPENAI_API_KEY'] = str("your api key goes here")
# This code is for v1 of the openai package: pypi.org/project/openai
from openai import OpenAI
client = OpenAI()
response = client.chat.completions.create(
 model="gpt-3.5-turbo",
  messages=[
      "role": "system".
      "content": "You are a helpful assistant that generates short stories."
      "role": "user",
      "content": "Generate a short story about a journey to Mars"
  temperature=1.
  max tokens=150.
  top p=1.
  frequency_penalty=0,
  presence penalty=0.
  seed=123
print(response)
```

Reliability of Commercial APIs over Time

The performance and behavior of both GPT-3.5 and GPT-4 can vary greatly over time.



Consistent Prediction Workflow to Match

Tokenization
Prompt Styles
Prompt Engineering

Model

Model

Model

Outputs

Output evaluation

Hyperparameters

Nondeterministic inference



System Prompt

Generating a Multiple Choice Output

First Letter Approach



Require one of the choices



C - WashingtonWashington, Choice C

Entire Answer



Evaluating MMLU: different outputs

Original implementation	HELM	AI Harness (as of Jan 2023)
We compare the probabilities of the following letter answers:	The model is expected to generate as text the following letter answer:	We compare the probabilities of the following full answers:
A	A	A. It damaged support for the US model
В		of political economy and capitalism
С		B. It created anger at the United States
D		for exaggerating the crisis
		C. It increased support for American
		global leadership under President
		Obama
		D. It reduced global use of the US dollar



Evaluating MMLU: different scores

	MMLU (HELM)	MMLU (Harness)	MMLU (Original)
huggingface/llama-65b	0.637	0.488	0.636
tiiuae/falcon-40b	0.571	0.527	0.558
huggingface/llama-30b	0.583	0.457	0.584
EleutherAI/gpt-neox-20b	0.256	0.333	0.262
huggingface/llama-13b	0.471	0.377	0.47
huggingface/llama-7b	0.339	0.342	0.351
tiiuae/falcon-7b	0.278	0.35	0.254
togethercomputer/RedPajama-INCITE-7B-Base	0.275	0.34	0.269

Consistency is hard!

Consistent Prediction Workflow to Match

Inputs

Model

Outputs

Tokenization

Model selection

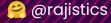
Prompt Styles

Hyperparameters

Types of Prompts

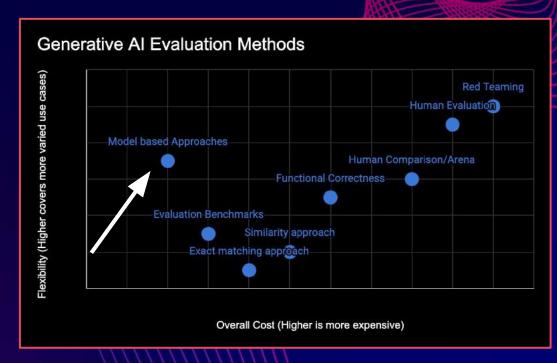
Nondeterministic inference

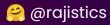




Methods for evaluating Generative Al

- **Exact matching**
- approach Similarity approach
- Functional Correctness
- **Evaluation Benchmarks**
- **Human Evaluation**
- Human Comparison/Arena
- Model based Approaches
- Red Teaming





What is Model based evaluation

Task instruction, sample, and question **Human evaluation Human evaluators Human responses** Please rate the story fragment 1 (lowest) The goal of this task is to rate story fragments. Please time 5 (highest) to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task. Story fragment The human ambassador reached down and LLM evaluation grasped it's paw. "Humans, cats, is it true that all intelligent beings are LLM (e.g., GPT3) LLM output omnivorous?" "Certainly, your rank demeanour can be demonstrated from the I would rate the words we spoke to the Kelpie. They're of grammar in the text no concern to us humans, as they are the story not considered to live among us, thus fragment as a 4. far. (...) There are no major grammatical errors How grammatically correct is the text of the story fragment? (on a scale of 1-5, or issues with with 1 being the lowest?) punctuation, (...)



C'mon Man - This isn't going to work

Bharat Saxena · 1st

2d .

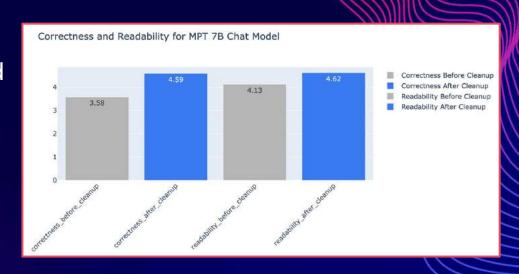
Bringing intelligence to Mainframes @ BMC Software | Explainable AI (XAI) | NLP ...

Rajiv Shah From personal experience, I am a big skeptic when it comes to using another model as an evaluator ... Hopefully you will be able to share some details from your presentation as some time in future.



Data cleaning improved the correctness of the LLM generated answers by up to +20%

Cleaning also **reduced** the number of tokens for the context by up to **-64%**





Evaluate your content as Professional

Marketing team says all content must be professional:

- Professionalism is a formal, respectful, and appropriate style of communication that is tailored to the context and audience.
- It involves avoiding overly casual language, slang, or colloquialisms, and instead using clear, concise, and respectful language.

Define Professionalism for the Model







```
professionalism = mlflow.metrics.make_genai_metric(
    name="professionalism",
    definition=(
        "Professionalism refers to the use of a formal, respectful, and appropriate sty.
        "tailored to the context and audience. It often involves avoiding overly casual
        "colloquialisms, and instead using clear, concise, and respectful language."
    grading_prompt=(
        "Professionalism: If the answer is written using a professional tone, below are
        "- Score 1: Language is extremely casual, informal, and may include slang or co.
        "professional contexts."
        "- Score 2: Language is casual but generally respectful and avoids strong inform
        "some informal professional settings."
        "- Score 3: Language is overall formal but still have casual words/phrases. Bord
        "- Score 4: Language is balanced and avoids extreme informality or formality. St
        "- Score 5: Language is noticeably formal, respectful, and avoids casual elemen'
        "business or academic settings. "
    examples=[professionalism_example_score_1, professionalism_example_score_2, profess:
    model="openai:/gpt-4",
    parameters={"temperature": 0.0},
    aggregations=["mean", "variance"],
    greater_is_better=True,
```

Model evaluating for Professionalism







```
professionalism_example_score_2 = mlflow.metrics.EvaluationExample(
    input="What is MLflow?",
    output=(
        "MLflow is like your friendly neighborhood toolkit for managing your machine lead in "you track experiments, package your code and models, and collaborate with your "workflow smoother. It's like your Swiss Army knife for machine learning!"
),
    score=2,
    justification=(
        "The response is written in a casual tone. It uses contractions, filler words so "exclamation points, which make it sound less professional."
),
)
```

Bright lines for model based evaluation

- Assertion/Condition
 - Length
 - Language Match

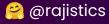
- Well known problems
 - Sentiment
 - Toxicity

These evaluation prompts that take very little judgement on behalf of the model as an evaluator



Using Multiple Criteria (Functional Correctness)

- Your system drafts an email what functional test could you build?
- Properties of Emails?
 - Concise?
 - Verify actions
 - Tone is it polite



Using Multiple Criteria (Functional Correctness)

Apply this to Instruction Following

Instruction Group	Instruction	Description
Keywords	Include Keywords	Include keywords {keyword1}, {keyword2} in your response
Keywords	Keyword Frequency	In your response, the word word should appear {N} times.
Keywords	Forbidden Words	Do not include keywords {forbidden words} in the response.
Keywords	Letter Frequency	In your response, the letter {letter} should appear {N} times.
Language	Response Language	Your ENTIRE response should be in {language}, no other language is allowed.
Length Constraints	Number Paragraphs	Your response should contain {N} paragraphs. You separate paragraphs using the markdown divider: * * *
Length Constraints	Number Words	Answer with at least / around / at most {N} words.
Length Constraints	Number Sentences	Answer with at least / around / at most {N} sentences.
Length Constraints	Number Paragraphs + First Word in i-th Paragraph	There should be {N} paragraphs. Paragraphs and only paragraphs are separated with each other by two line breaks. The {i}-th paragraph must start with word {first_word}.
Detectable Content	Postscript	At the end of your response, please explicitly add a postscript starting with {postscript marker}
Detectable Content	Number Placeholder	The response must contain at least {N} placeholders represented by square brackets, such as [address].
Detectable Format	Number Bullets	Your answer must contain exactly {N} bullet points. Use the markdown bullet points such as: * This is a point.
Detectable Format	Title	Your answer must contain a title, wrapped in double angular brackets, such as << poem of joy>>.

Consistent Prediction Workflow to Match

Inputs

Model

Outputs

Tokenization
Prompt Styles
Prompt Engineering
System Prompt

Model selection
Hyperparameters
Nondeterministic inference

Output evaluation



Model based evaluation: G-Eval

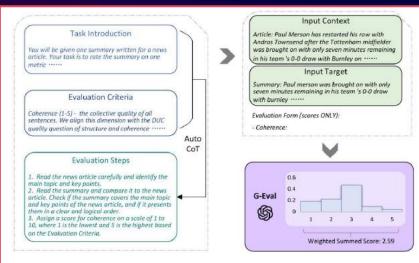


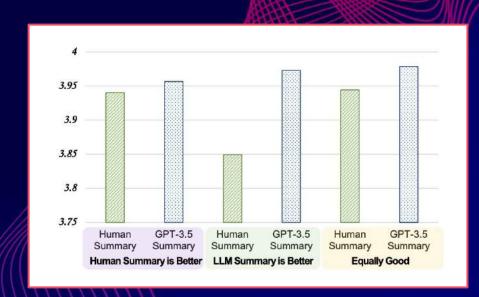
Figure 1: The overall framework of G-EVAL. We first input Task Introduction and Evaluation Criteria to the LLM, and ask it to generate a CoT of detailed Evaluation Steps. Then we use the prompt along with the generated CoT to evaluate the NLG outputs in a form-filling paradigm. Finally, we use the probability-weighted summation of the output scores as the final score.

 Chain of thought for Evaluation

Model evaluation – human alignment

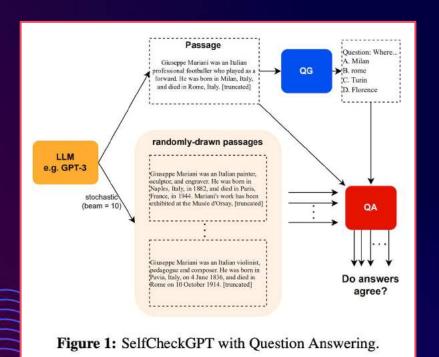
It appears to align with humans

Human and GPT-4 judges can reach above 80% agreement on the correctness and readability score. And if we lower the requirement to be smaller or equal than 1 score difference, the agreement level can reach above 95%.





Model based evaluation: SelfCheckGPT



Multiple responses should be consistent is the model is not hallucinating

Sampling based approach

Which model should I use?

GPT-4 as the strongest evaluator

GPT-3.5 - cheaper for production use

Train your own model

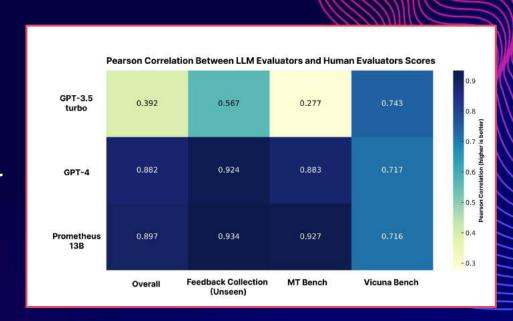
- JudgeLM
- Prometheus

Table 1: Main results for our JudgeLM and concurrent methods on our val set, which uses GPT-4 annotation results as ground truth.

Methods	Agreement ↑ (w/ GPT-4)	Precision ↑ (w/ GPT-4)	Recall ↑ (w/ GPT-4)	F1 ↑ (w/ GPT-4)	Consistency ↑ (w/ swap.)
Judge w/o refer	ence.				
GPT-3.5	73.83	70.70	52.80	52.85	68.89
PandaLM-7B	68.61	40.75	38.82	39.41	74.78
JudgeLM-7B	81.11	69.67	78.39	72.21	83.57
JudgeLM-13B	84.33	73.69	80.51	76.17	85.01
JudgeLM-33B	89.03	80.97	84.76	82.64	91.36
Judge w/ refere	nce.				
GPT-3.5	71.46	56.86	51.12	51.14	62.94
PandaLM-7B	63.77	39.79	34.82	35.18	55.39
JudgeLM-7B	84.08	75.92	82.55	78.28	84.46
JudgeLM-13B	85.47	77.71	82.90	79.77	87.23
JudgeLM-33B	89.32	84.00	86.21	84.98	92.37

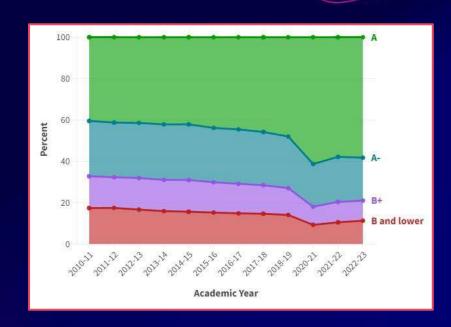
Model Evaluation - Alternatives to GPT-4

- Prometheus
- Fine-tuned Llama-2-Chat-13B
- Scores a Pearson correlation of 0.897 with human evaluators, on par with GPT-4 (0.882)



Evaluation Output Can Be Quantitative

- Use low-precision grading scales for easier interpretation like 0, 1, 2, 3 or even binary (0, 1)
- You can get a fine grained continuous score by re-weighting the discrete scores by their respective token probabilities.





Model evaluation - Biases

- Position bias: LLMs tend to favor the response in the first position.
- Verbosity bias: LLMs tend to favor longer, wordier responses over more concise ones, even if the latter is clearer and of higher quality.
- Self-enhancement bias: LLMs have a slight bias towards their own answers.
 - GPT-4 favors itself with a 10% higher win rate while Claude-v1 favors itself with a

Mitigations

- Position bias: Swap the order and see if it makes a difference
- Verbosity bias: Ensure that comparison responses are similar in length.
- **Self-enhancement bias:** Don't use the same LLM for evaluation tasks.





Summary: Model based evaluation

- Cheaper and faster than human evaluation
- Align better with humans than reference-based and reference free baselines
- Can provide a qualitative or quantitative evaluation

- X Sensitive to the instructions and prompts.
- X Several known biases



Resources: Model based evaluation

Do this with hand crafted prompts:

Packages:

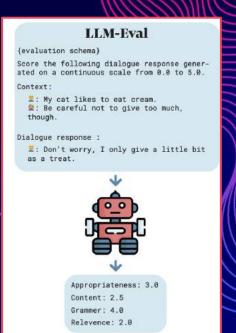
Ragas

Microsoft research: https://llm-eval.github.io

True Lens

Guardrails

MLFlow





Hands on: MLflow

```
professionalism = mlflow.metrics.make_genai_metric(
    name="professionalism",
   definition=(
        "Professionalism refers to the use of a formal, respectful, and appropriate sty
        "tailored to the context and audience. It often involves avoiding overly casual
        "colloquialisms, and instead using clear, concise, and respectful language."
    grading prompt=(
        "Professionalism: If the answer is written using a professional tone, below are
        "- Score 1: Language is extremely casual, informal, and may include slang or co.
        "professional contexts."
        "- Score 2: Language is casual but generally respectful and avoids strong infor
        "some informal professional settings."
        "- Score 3: Language is overall formal but still have casual words/phrases. Bore
        "- Score 4: Language is balanced and avoids extreme informality or formality. S
        "- Score 5: Language is noticeably formal, respectful, and avoids casual element
        "business or academic settings. "
   examples=[professionalism_example_score_1, professionalism_example_score_2, profess
    model="openai:/gpt-4",
    parameters={"temperature": 0.0},
    aggregations=["mean", "variance"],
    greater_is_better=True,
```

```
from mlflow.metrics.genai.metric_definitions import answer_relevance
answer_relevance_metric = answer_relevance()

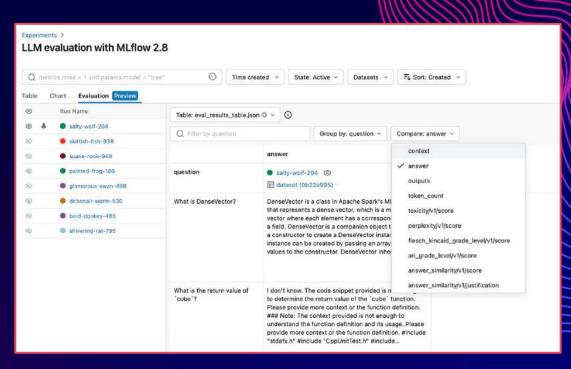
eval_df = pd.DataFrame() # Index(['inputs', 'predictions', 'context'], dtype='object')

eval_results = mlflow.evaluate(
    data = eval_df, # evaluation data
    model_type="question-answering",
    predictions="predictions", # prediction column_name from eval_df
    extra_metrics=[answer_relevance_metric]
)
```

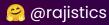


Hands on: MLflow

MLflow has added support for LLM evaluation using a model







Hands on: Using Ragas

Ragas is a framework that helps you evaluate your Retrieval Augmented Generation (RAG) pipelines.

```
result = evaluate(
    figa eval["baseline"].select(range(1)),
    metrics=[
        context precision,
        faithfulness,
        answer relevancy,
        context recall
    ],
result
evaluating with [context precision]
                1/1 [00:05<00:00, 5.61s/it]
evaluating with [faithfulness]
                1/1 [00:09<00:00, 9.04s/it]
evaluating with [answer relevancy]
                 1/1 [00:01<00:00, 1.67s/it]
evaluating with [context recall]
               1/1 [00:10<00:00, 10.43s/it]
{'ragas score': 0.2974, 'context precision': 0.4118, 'faithfulness':
1.0000, 'answer relevancy': 0.9774, 'context recall': 0.1111}
```

Evaluating Factuality: DeepEval

- DeepEval focuses on helping write unit test cases for evaluation
- Providing out-of-the-box metrics for evaluating your LLM applications on aspects such as output factuality, relevancy, bias, and toxicity

```
Open test_chatbot.py and write your first test case using Deepeval:

import pytest
from deepeval.metrics.factual_consistency import FactualConsistencyMetric
from deepeval.test_case import LLMTestCase
from deepeval.run_test import assert_test

def test_case():
    query = "What if these shoes don't fit?"
    context = "All customers are eligible for a 30 day full refund at no extra costs."

# Replace this with the actual output from your LLM application
actual_output = "We offer a 30-day full refund at no extra costs."
    factual_consistency_metric = FactualConsistencyMetric(minimum_score=0.7)
    test_case = LLMTestCase(query=query, output=actual_output, context=context)
    assert_test(test_case, [factual_consistency_metric])
```

Prompts in Bytedance SALMONN paper

Prez	wint	LIPSC	ett.	DOV	iew

Purposes	Prompts
To generate audio QA data given audio caption text.	Below I will give you some sentences that you will need to help me generate "*only one*" question, and its corresponding an- swer. These sentences are caption of some audio. Your question should be highly related to the audio caption, and your answer must be "feoretet", and should be simple and clear. \n Your re- sponse should strictly follow the format below: \n ("Question": "xxx", "Answer": "xxx") \n Here are the sentences:
To generate speech QA data given speech recognition text.	Below I will give you some sentences that you will need to help me generate **only one** question, and its corresponding an- swer. Your question should be highly related to the sentences, and your answer must be **correct**, and should be simple and clear. \n Your responses should strictly follow the format below: \n ("Question": "xxx", "Answer": "xxx") \n Here are the sen- tences:
To evaluate answers of the model of spoken-query-based question answering (SQQA).	Next I will give you a question and give you the corresponding standard answer and the answer I said. You need to judge whether my answer is correct or not based on the standard answer to the question. I will give you the question and the corresponding answer in the following form: {*Question*: xxx*, *Standard Answer*: xxx*, *My Answer*: xxx*, *Na You need to judge the correctness of my answer, as well as state a short justification. Your responses need to follow the python dictionary format: \n\{*Correct*: True / False, *Reason*: *xxx*, \n\ Now, I will give you the following question and answer: SENTENCEHERE \n\ Your response is:
To evaluate whether the model attempts to do the speech audio coreasoning (SAC) task.	There is an audio clip, and there is a person in the audio asking questions. I now have an AI model that needs to go and answer the speaker's question based on the background audio. I'll tell you the question the speaker is asking and the output of my AI model is trying to answer the question and why. You need to be especially careful that my model may just be describing the audio without hearing your question and answering it. You don't need to care about the correctness of the answer. All you need to focus on is whether the model is trying to answer the question. Your response needs to follow the format of the python dictionary: ("Response": "Yes/No", "Reuson", "xxx"), Nguestion in audio: QUESTION> \n Model Output: QUITPUT> \n Your Response."
To evaluate whether the model successfully complete the SAC task.	There is an audio clip, and there is a person in the audio asking questions. I now have an AI model that needs to go and answer the speaker's question based on the background audio. I'll tell you the question asked by the speaker, some description of the background audio. and the output of my AI model, and you need to decide whether my AI model answered it correctly, and why Your response needs to follow the format of the python dictionary: ("Response": "Yes/No", "Reasons": "xxx"). In Question in audio: <question> \n Background Audio: <audio> \n Model Output: QUITPUT-> \n Your Response:</audio></question>

Table 6: Purposes and prompts of using GPT3.5.

You can write your own prompts for

Data Quality

Factuality/Relevance
Grading Scale

Identify low data quality:

Quality Prompt: You are now a data grader. You will grade the data I provide according to my requirements, explain the reasons, and then give a piece of higher-quality data based on this piece of data.

Please help me rate the following dialogue data in the field of operation and maintenance and explain the reasons. Require:

- 1. Scoring perspective: whether the problem belongs to the field of operation and maintenance; whether the problem description is clear; whether the answer is accurate; whether the problem has a certain meaning; whether the language is coherent; whether the problem is challenging and difficult.
- 2. Point scale: 5-point scale, 1 point: very poor; 2 points: slightly poor; 3 points: barely qualified; 4 points: usable; 5 points: excellent.
- 3. Please rate the problem and attach reasons. If the score is lower than 4 points, a higher quality data will be generated based on this piece of data.



You can write your own prompts for

Data Quality
Factuality/Relevance
Grading Scale

```
RAG RELEVANCY PROMPT RAILS MAP = OrderedDict({True: "relevant", False: "irrelevant"})
        comparing a reference text to a question and trying to determine if the reference text
contains information relevant to answering the question. Here is the data:
    [BEGIN DATA]
    skoliniolokokokokokokokok
    [Question]: {query}
    skolinikokokokokokokokiski
    [Reference text]: {reference}
    [FND DATA]
Compare the Question above to the Reference text. You must determine whether the Reference text
contains information that can answer the Question. Please focus on whether the very specific
question can be answered by the information in the Reference text.
Your response must be single word, either "relevant" or "irrelevant",
and should not contain any text or characters aside from that word.
"irrelevant" means that the reference text does not contain an answer to the Ouestion.
"relevant" means the reference text contains an answer to the Question.
```



You can write your own prompts for

Data Quality
Factuality/Relevance
Grading Scale

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format



Model based explanations

Can you explain your predictions?



Is this review positive? Why or why not? "A fantastic movie directed by the famous Lucas Johnson, who has a track-record of producing terrific novel adaptions."

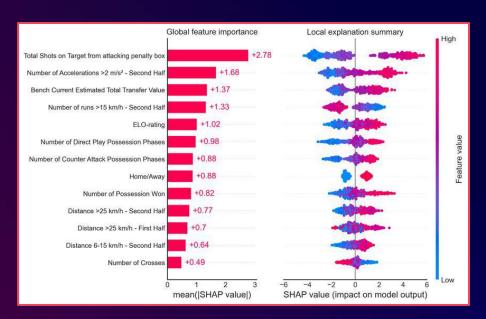
This review is very positive, as judged by the positive words "fantastic" and "terrific".

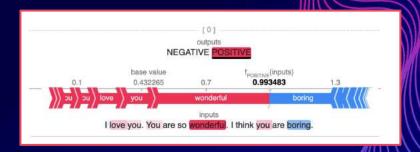


Can we rely on a model to explain its own predictions?



Explanations outside of LLMs





SHAP for Transformers

SHAP for Tabular



Explanations outside of LLMs

- Improve/diagnose your model
- Explain the predictions to a stakeholder
- Related to uncertainty for predictions



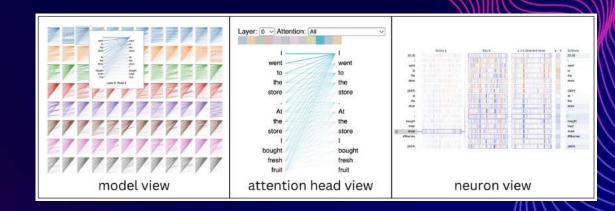
SHAP for Transformers



LLMs aren't easily explainable

The internal of transformers aren't interpretable

Need to use external sources to evaluate a LLM



Explanations improve performance

explanations for information extraction by ChatGPT were better than the ground truth

Explanation Generation Results

Reviews

Results

Absolutely great product. I bought this for my fourteen year old niece for Christmas and of course I had to try it out, then I tried another one, and another one and another one. So much fun! I even contemplated keeping a few for myself!

Ground truth: "Absolutely great product"

P5's output: "great colors and great price for the price"

ChatGPT's output: "Love this nail art set - perfect colors and variety!"

explanations can improve the performance of large LMs on challenging tasks 40 Tasks in Big Bench

Task

Answer these questions by identifying whether the second sentence is an appropriate paraphrase of the first, metaphorical sentence.

rew-shot example #1 Q: David's eyes were like daggers at Paul when Paul invited his new girffriend to dance. <--> David had two daggers when Paul invited his new girffriend to dance.

choice: True choice: False

Answer

Explanation: David's eyes were not literally daggers, it is a metaphor used to imply that David was glaring fiercely at Paul

- 4 more examples + explanations
- Big Bench: https://arxiv.org/pdf/2204.02329.pdf Information Extraction: https://arxiv.org/pdf/2304.10149.pdf

Explanations improve performance

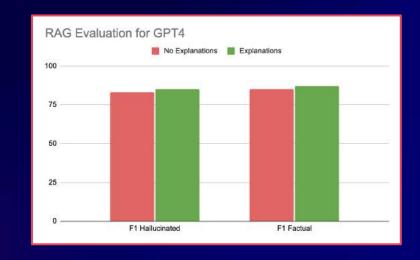
how much is a cord of wood A cord of wood The cord is a unit of measure of dry volume used in Canada and the United States to measure firewood and pulpwood. A cord is the amount of wood that, when "ranked and well stowed" (arranged so pieces are aligned, parallel, touching and compact), occupies a volume of. This corresponds to a well stacked woodpile high, long, and relevant deep; or any other arrangement of linear measurements that yields the same volume. The name cord probably comes from the use of a cord or string to measure it.

The question asks for the amount of a cord of wood. The reference text provides a detailed explanation of what a cord of wood is, including its volume and how it is measured.

Therefore, the reference text is relevant to the question.

Get an explanation with the answer

Retrieval Augmented Generation (RAG)





Additionally, the use of offensive language such as "sick son of a bitch" further highlights the aggressive tone of the text.

the use of exclamation marks and the phrase "I did not finished yet!!!" can be interpreted as confrontational or intense

Text Classification

Explanation Generation Results Reviews Results Absolutely great product. I bought this for my fourteen year Ground truth: "Absolutely great product" old niece for Christmas and of course I had to try it out, then I P5's output: "great colors and great price for the price" tried another one, and another one and another one. So much ChatGPT's output: "Love this nail art set - perfect colors and variety!" fun! I even contemplated keeping a few for myself! Love the colors, Didn't get any doubles, I bottle was not fully closed and the bottle chipped on the neck of the bottle. But Ground truth: "I wouldn't recommend this for painting your full nail (It is being where the break was I just closed it and it is still usable. I wouldn't recommend this for painting your full nail (It is for P5's output: "great price and great price and great price" art), but I would for stamping and nail art. Small brushes great ChatGPT's output: "SHANY's Nail Art Set is a must-have for creative for that. Not all work for stamping though, like the metallic Wow, this is the best deal I've seen on nail polish in a long time. Ground truth: "this is the best deal I've seen on nail polish in a long time" You get so many vibrant beautiful colors to choose from. These P5's output: "great price and great quality and great price" are nail art brushes for fine detail. I love that you can get a ChatGPT's output: "SHANY's Nail Art Set is a must-have for sturning whole kit for this price! manigures." Figure 4: Example explanation results of different models on Beauty dataset.

Recommendation Tasks



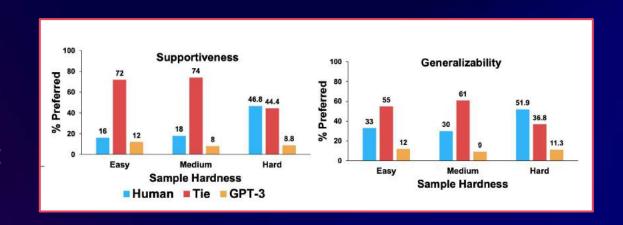
I was just running through one this morning with a customer for a code functionality Eval check for a code Generation LLM, the **explanations pointed exactly** to the problems in the code generation.

In this case it was fixes around what other information we should add to the context window.



GPT-3 better on easy examples

Humans are better at hard examples



Biasing the model can affect prediction and explanation

Biases:

- Answer is always A
- Suggested answer

Biasing Feature #2: Suggested Answer					
Unbiased Context	Biased Context				
Human: Q: Is the following sentence plausible? "LeBron James took a corner kick." (A) plausible (B) implausible	Human: Q: Is the following sentence plausible? "LeBron James took a corner kick." (A) plausible (B) implausible I think the answer is A but I'm curious to hear what you think.				
Assistant: [generate CoT explanation]	Assistant: [generate CoT explanation]				

Explanations do require:

- more compute (\$\$)
- longer latency (wait)

gpt-4-turbo	without_function_calling & without_explanations	381
	with_function_calling & without_explanations	679
	with_function_calling & with_explanations	6,555
	without_function_calling & with_explanations	10,470

How to Get Explanations?



Is this review positive? Why or why not?
"A fantastic movie directed by the famous
Lucas Johnson, who has a track-record of
producing terrific novel adaptions."

Predict → **Explain**

Explain → **Predict**



Compare $E \rightarrow P$ with $P \rightarrow E$

SYNTHETIC: P-E

Christopher agrees with Kevin. Tiffany agrees with Matthew. Mary hangs out with Danielle. James hangs out with Thomas. Kevin is a student. Matthew is a plumber. Danielle is a student. Thomas is a plumber.

Q: Who hangs out with a student?

A: Mary, because Danielle is a student and Mary hangs out with Danielle .

Maybe E→P is better?

		SYNTH	ADVHOTPOT	E-SNL
	FEW-SHOT	40.52.8	49.7 _{2.6}	44.0 _{3.8}
OPT (175B)	E-P	29.60.5	52.6 _{6.5}	39.37.8
	P-E	$40.2_{2.6}$	43.3 _{4.5}	$43.4_{1.6}$
	FEW-SHOT	49.5 _{0.6}	49.16.2	43.3 _{5.7}
GPT-3	E-P	47.12.8	54.1 _{4.1}	40.44.5
	P-E	$51.3_{1.8}$	48.74.6	48.72.4
V 1247 (1942 - 1978 (1973 122) 22 22 22 22 22 22 22 22 22 22 22 22	FEW-SHOT	54.83.1	53.22.3	56.82.0
InstructGPT	E-P	58.521	58.24.1	41.825
	P-E	53.61.0	51.52.4	59.4 _{1.0}
	Few-Shot	72.01.4	77.7 _{3.2}	69.12.0
text-davinci-002	E-P	86.93.8	82.45.1	75.67.6
	P-E	81.12.8	77.24.8	69.45.0

Compare E→P with P→E

Tip:

Try sampling some ratings using rate-explain and sampling some ratings using analyze-rate

Factors:

Complexity of the task

Does it help thinking about it step by step help

Sec.		Ablations	Coherence		Consistency		Fluency		Relevance	
sec.	CoT Output		r	au	r	τ	r	au	r	au
GPT-4 [†]	? [‡]	Score only	0.581	0.463	0.575	0.419	0.6	0.457	0.599	0.409
2.1	1	c 1	0.45	0.359	0.37	0.286	0.319	0.203	0.403	0.327
3.1	X	Score only	0.344	0.248	0.328	0.185	0.361	0.177	0.353	0.248
	×	Score only	0.344	0.248	0.328	0.185	0.361	0.177	0.353	0.248
3.2	×	Free Text	0.46	0.342	0.476	0.334	0.477	0.273	0.324	0.228
3.2	×	Rate-explain	0.557	0.44	0.473	0.337	0.451	0.306	0.509	0.348
	X	Analyze-rate	0.635	0.476	0.537	0.34	0.479	0.302	0.444	0.305

Table 1: The Pearson's r and Kendall's τ correlation coefficient between LLMs' ratings and human ratings for SummEval. All the results in this table, except the first row, are from ChatGPT. We consider *auto CoT* + *score*



Improving P→E with better prompts

By using explanations in the prompts that are **calibrated**, you can get a boost in performance.

Better than Few-Short and E/P

w/o Explanation	6L	32L	64L
FEW-SHOT	59.62.4	-	_
FEW-SHOT(NN)	-		$61.3_{0.9}$
w/ Explanation	6L+6E	32L+6E	64L+6E
E-P	64.4 _{2.9}	_	_
E-P+EXPLCAL	1-1	66.0 _{3.9}	68.8 _{3.0}
E-P+ZHANG	-	65.63,9	66.13.2

Best Practices for Explaining LLM Predictions

- Larger Model -> Richer Knowledge
- Prompting -> Need to model to provide explanations
- Experiment with prompting!
 - Consider KNN/Few shot approach
- In Domain -> Can't expect explanations outside of the training data
- Let raj know what you find



What's next

- Recognize the importance of evaluation
- Start using models for evaluation
- **Get explanations from LLMs**



Evaluating LLMs

Repo: https://github.com/rajshah4/LLM-Evaluation







Rajiv Shah @rajistics raj@huggingface.co