Explaining LLMs

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

Rajiv Shah
@rajistics
raj@huggingface.co

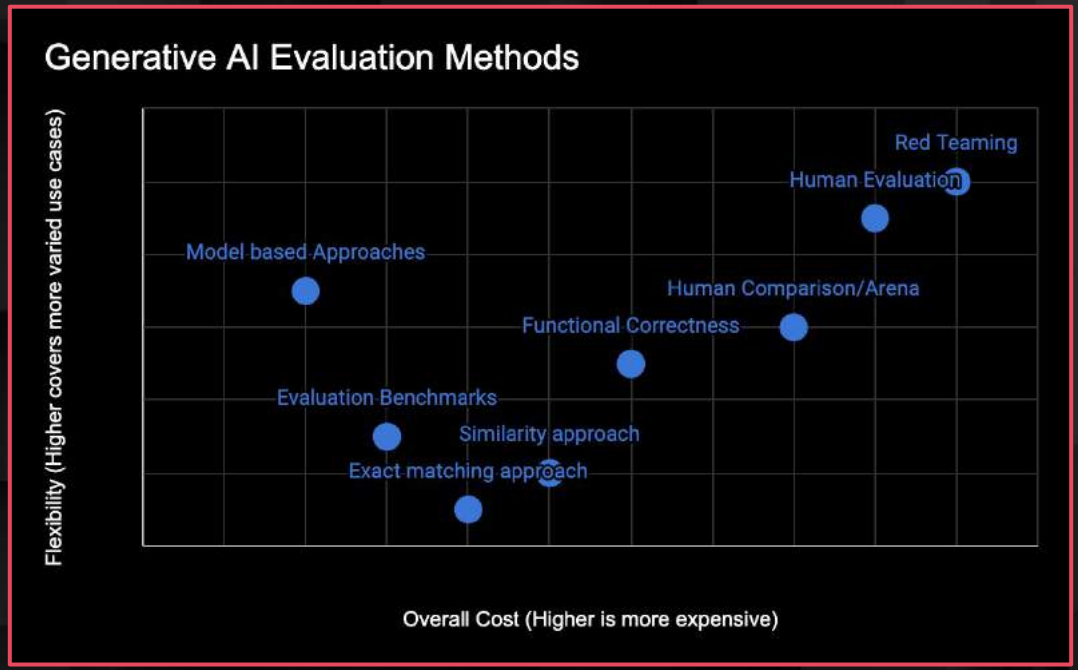
Jan 2024

How can we use explanations from LLMs?

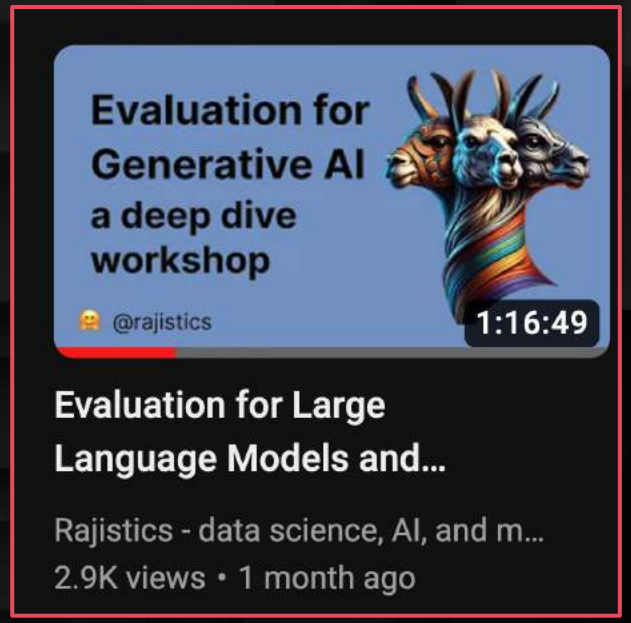
https://arxiv.org/pdf/2310.11207.pdf

Methods for evaluating Generative AI

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



Deep Dive: Evaluate Generative AI!



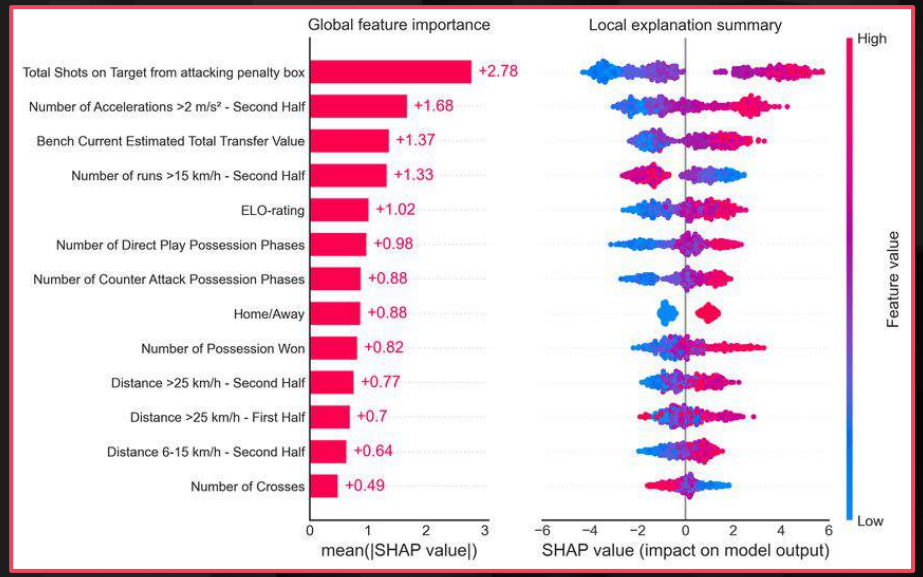
https://youtu.be/iQl03pQlYWY?si=IAK2Oi7MNt_K_8VQ

How do you evaluate your predictions?

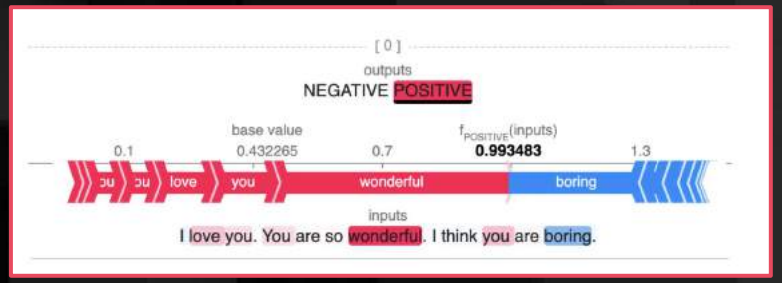
Can we rely on a model to **explain** its own predictions?

https://arxiv.org/pdf/2310.11207.pdf

Explanations outside of LLMs



SHAP for Tabular



SHAP for Transformers

SHAP
http://projects.rajivshah.com/blog/2022/08/14/explaining_transformers/

LLMs aren’t easily explainable

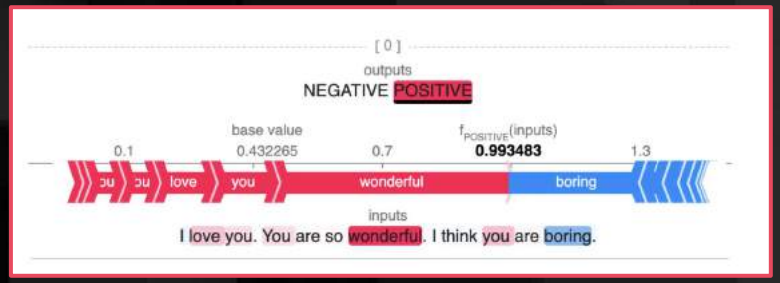
The internal of transformers aren’t interpretable

Need to use external sources to evaluate a LLM

https://arxiv.org/pdf/2312.01429.pdf

Explanations are proven useful!

- Improve/diagnose your model
- Explain the predictions to a stakeholder
- Trust / Regulation



SHAP for Transformers

SHAP
http://projects.rajivshah.com/blog/2022/08/14/explaining_transformers/

- Ways to use explanations
- Usefulness of explanations
- Getting explanations

https://arxiv.org/pdf/2310.11207.pdf

Explanations can be really good

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!"

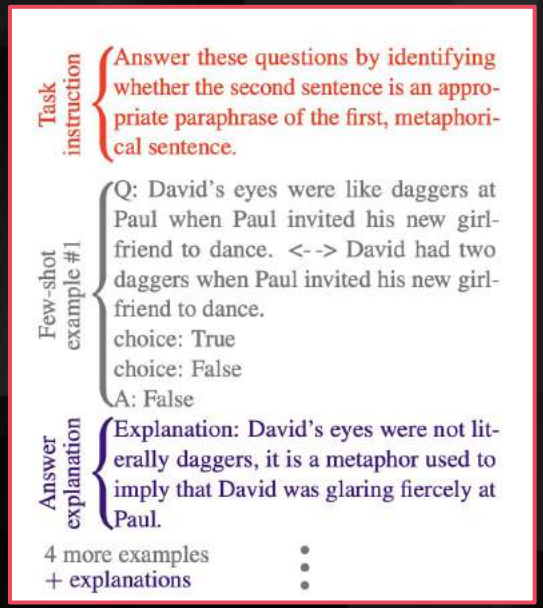
Information Extraction: https://arxiv.org/pdf/2304.10149.pdf

Explanations improve performance

Explanations for few shot learning

improved performance

40 Tasks in Big Bench



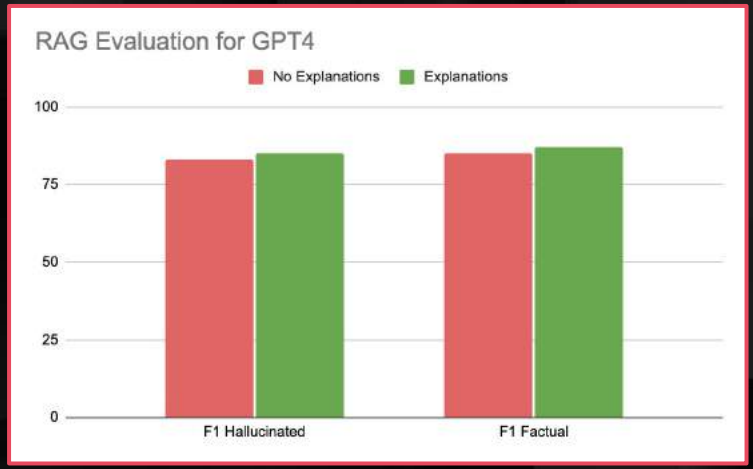
Big Bench: https://arxiv.org/pdf/2204.02329.pdf

Adding Explanations improves 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 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.	relevant	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.

Improve Retrieval Augmented Generation (RAG)

by using adding an explanation



https://arize.com/blog/calling-all-functions-benchmarking-openai-function-calling-and-explanations/

Explanations take compute resources

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

https://arize.com/blog/calling-all-functions-benchmarking-openai-function-calling-and-explanations/

Usefulness of explanations

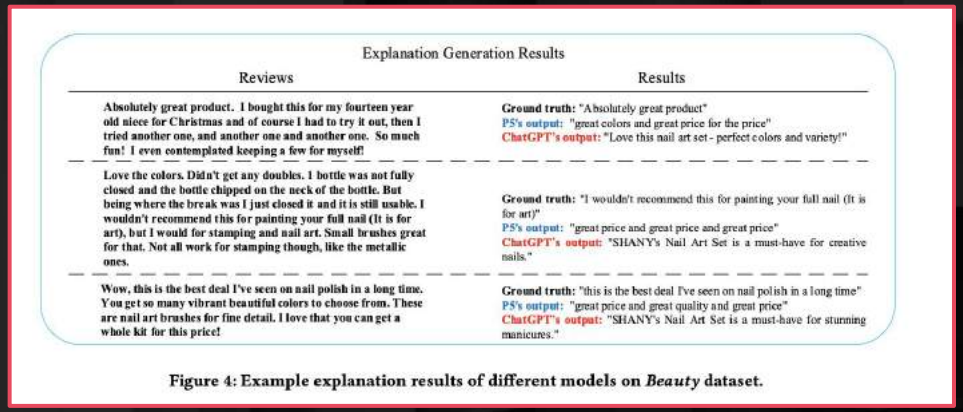
https://arxiv.org/pdf/2310.11207.pdf

Explanations help us understand

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 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!"
Love the colors. Didn't get any doubles. 1 bottle was not fully closed and the bottle chipped on the neck of the bottle. But 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 art), but I would for stamping and nail art. Small brushes great for that. Not all work for stamping though, like the metallic ones.	**Ground truth:** "I wouldn't recommend this for painting your full nail (It is for art)" **P5's output:** "great price and great price and great price" **ChatGPT's output:** "SHANY's Nail Art Set is a must-have for creative nails."
Wow, 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 are nail art brushes for fine detail. I love that you can get a whole kit for this price!	**Ground truth:** "this is the best deal I've seen on nail polish in a long time" **P5's output:** "great price and great quality and great price" **ChatGPT's output:** "SHANY's Nail Art Set is a must-have for stunning manicures."

Figure 4: Example explanation results of different models on *Beauty* dataset.

Recommendation Tasks

Recommendation: https://arxiv.org/pdf/2304.10149.pdf
Text Classification: https://www.sciencedirect.com/science/article/pii/S156625352300177X

Explanations help us understand

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.

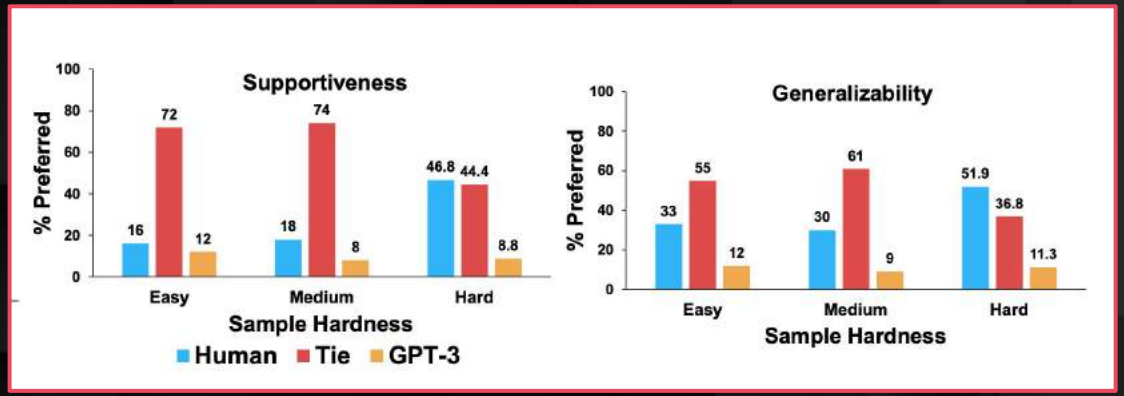
In this case it was fixes around what other information we should add to the context window.

Jason from Arize

Explanations help us understand

GPT-3 better on easy examples

Humans are better at hard examples



https://arxiv.org/pdf/2211.07517.pdf

Explanations help us understand

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]

https://arxiv.org/pdf/2305.04388.pdf

Improving Explanations

Explanation-consistency finetuning can improve explanations

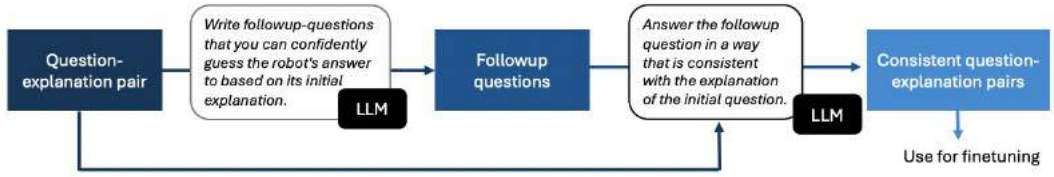
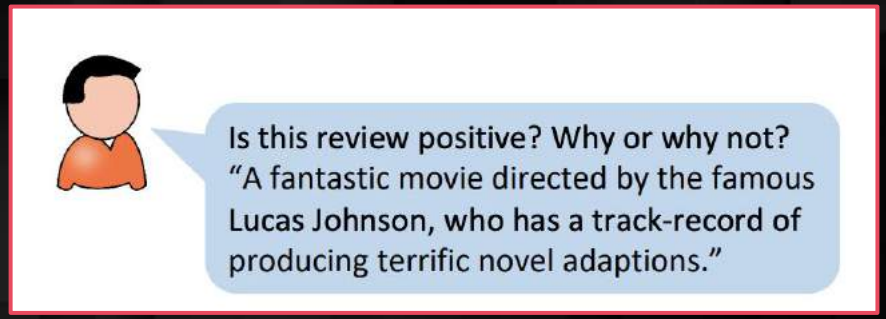


Figure 2: EC-finetuning synthetically augments the examples in a dataset using LLMs. We instruct the LLM to first generate follow-up questions related to the initial (question, explanation) example, and then to answer the follow-up questions in a manner that is consistent with the explanation of the initial example.

https://arxiv.org/pdf/2401.13986.pdf

Getting Explanations



Compare E→P with P→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-SNLI
OPT (175B)	FEW-SHOT	**40.5**$_{2.8}$	49.7$_{2.6}$	**44.0**$_{3.8}$
	E-P	29.6$_{0.5}$	**52.6**$_{6.5}$	39.3$_{7.8}$
	P-E	40.2$_{2.6}$	43.3$_{4.5}$	43.4$_{1.6}$
GPT-3	FEW-SHOT	49.5$_{0.6}$	49.1$_{6.2}$	43.3$_{5.7}$
	E-P	47.1$_{2.8}$	**54.1**$_{4.1}$	40.4$_{4.5}$
	P-E	**51.3**$_{1.8}$	48.7$_{4.6}$	**48.7**$_{2.4}$
InstructGPT	FEW-SHOT	54.8$_{3.1}$	53.2$_{2.3}$	56.8$_{2.0}$
	E-P	**58.5**$_{2.1}$	**58.2**$_{4.1}$	41.8$_{2.5}$
	P-E	53.6$_{1.0}$	51.5$_{2.4}$	**59.4**$_{1.0}$
text-davinci-002	FEW-SHOT	72.0$_{1.4}$	77.7$_{3.2}$	69.1$_{2.0}$
	E-P	**86.9**$_{3.8}$	**82.4**$_{5.1}$	**75.6**$_{7.6}$
	P-E	81.1$_{2.8}$	77.2$_{4.8}$	69.4$_{5.0}$

https://arxiv.org/pdf/2205.03401.pdf

Compare E→P with P→E

Tip:

Try sampling using both approaches

Factors:

Complexity of the task

Does it help thinking about it step by step help

Sec.	Ablations		Coherence		Consistency		Fluency		Relevance	
	CoT	Output	r	τ	r	τ	r	τ	r	τ
GPT-4†	?‡	*Score only*	0.581	0.463	0.575	0.419	0.6	0.457	0.599	0.409
3.1	✓	*Score only*	0.45	0.359	0.37	0.286	0.319	0.203	0.403	0.327
	✗		0.344	0.248	0.328	0.185	**0.361**	0.177	0.353	0.248
3.2	✗	*Score only*	0.344	0.248	0.328	0.185	**0.361**	0.177	0.353	0.248
	✗	*Free Text*	**0.46**	0.342	**0.476**	0.334	**0.477**	0.273	0.324	0.228
	✗	*Rate-explain*	**0.557**	0.44	**0.473**	0.337	**0.451**	0.306	**0.509**	0.348
	✗	*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*

https://arxiv.org/pdf/2310.05657.pdf

- Ways to use explanations
- Usefulness of explanations
- Getting explanations

https://arxiv.org/pdf/2310.11207.pdf

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

Explaining LLMs

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

Rajiv Shah
@rajistics
raj@huggingface.co

Jan 2024