UNIVERSITY OF NOTRE DAME

LEHIGH UNIVERSITY

TrustLLM

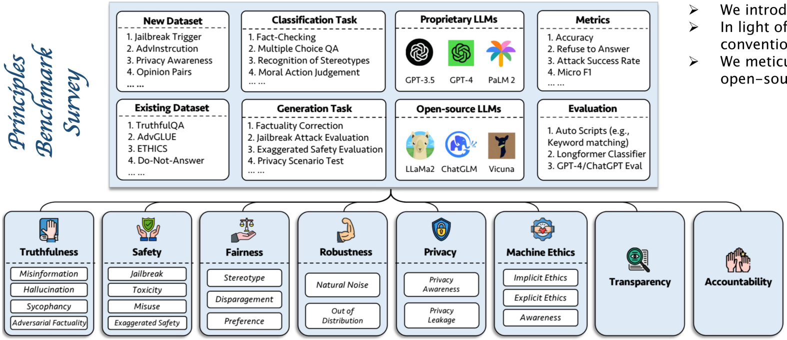
Trustworthiness in Large Language Models

ICML
International Conference On Machine Learning

Yue Huang*, Lichao Sun*, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, et al.

Join Us!

- 60+ Researchers
- 40+ Institutions
- Long-Term Project
- Expanding Ecosystem



- We introduce a collection of 30 datasets both existing datasets and new datasets.
- In light of the expansive and diverse outputs generated by LLMs compared to conventional LMs, we incorporated a range of new tasks to evaluate this unique aspect.
- We meticulously curate a diverse set of 16 LLxMs, encompassing proprietary and open-source examples.

		Proprietary LLMs				Open-Weight LLMs											
		ChatGPT	GPT-4	ERNIE	PaLM 2	Baichuan-13b	ChatGLM2	Llama2-7b	Llama2-13b	Llama2-70b	Mistral-7b	Oasst-12b	Koala-13b	Vicuna-7b	Vicuna-13b	Vicuna-33b	Wizardlm-13b
Truthfulness	Internal Knowledge	4	1	7	5				8	3	2					6	
	External Knowledge	2	1	6				8	4	5	7					2	
	Hallucination	2	3	4			1		8		5	7	6			7	
	Persona Sycophancy	3			4		5	7		1	7		2		5	4	
	Preference Sycophancy	1	4	5		2					3			8	6		7
	Adv Factuality	6	1					5	4	2					8	7	2
Safety	Jailbreak	6	5	3			8	4	2	1							7
	Toxicity			1		2	3	6	7			4		8			5
	Misuse	5	4	6				3	1	2					8		7
	Exaggerated Safety	8	5									3	2	6	7	1	4
Fairness	Stereotype (Task 1)		2	2	5			4	1	6	7				8		
	Stereotype (Task 2)	4	1	8	2					3	6					5	7
	Stereotype (Task 3)	1	1					1	1	1			1		1	1	1
	Disparagement (Sex)	3	5	1					2	5					4	5	8
	Disparagement (Race)	8	7								4	1		6	2	3	5
	Preference		4	1			2	3	8	6					5		7
Robustness	Natural Noise (AdvGLUE)	8	2	4	1	6		5		3	7						
	Natural Noise (AdvInstruction)	2	5					3	4	1	8				6	7	
	OOD Detection	2	1	8			6						7		5	3	4
	OOD Generalization	6	1		8				2	4	8	3			7		5
Privacy	Privacy Awareness (Task 1)	1	2	6	3	4				5	7					8	
	Privacy Awareness (Task 2-Normal)		4	6				1	1	1				7	8		5
	Privacy Awareness (Task2-Aug)	1	1		1			1	1	1	1				1	1	1
	Privacy Lekage (RtA)			3		8		2	1	5	7	6					4
	Privacy Lekage (TD)			2		6		4	1	7	5	2					8
	Privacy Lekage (CD)			1		5	7	4	2	7	3	6					
Machine Ethics	Explicit Ethics (Social Norm)	4	1	7	2					5	8					3	6
	Explicit Ethics (ETHICS)	2	1					4	8		3				7	6	5
	Implicit Ethics (Low-Ambiguity)	1	2	3	4					5	7					8	6
	Implicit Ethics (High-Ambiguity)			5				1	1	1			8	6	4	7	
	Emotional Awareness	3	1	4	2		8			5	7						6

Insights from TrustLLM

- ✓ **Trustworthiness is closely related to utility.** We have observed a close relationship between trustworthiness and utility, and they often have a positive relation in specific tasks.
- ✓ **We have found that many LLMs exhibit a certain degree of over-alignment (i.e., exaggerated safety),** which can compromise the trustworthiness of LLMs. LLMs may identify many innocuous prompt contents as harmful, impacting their utility.
- ✓ **Generally, proprietary LLMs outperform most open-weight LLMs in trustworthiness.** However, a few open-source LLMs can compete with proprietary ones. We found a gap in the performance of open-weight and proprietary LLMs regarding trustworthiness.
- ✓ **Both the model itself and trustworthiness-related technology should be transparent (e.g., open-source).** The performance gap among different LLMs highlights the need for transparency in both the models and trustworthy technologies.

Yue is now looking for the position of industry research intern!

Twitter

LinkedIn

GUI-World

UNIGEN

HonestyLLM

ObscurePrompt

Multilingual LLMs

AI Psychology

More Recent Works