NEURAL INFORMATION PROCESSING SYSTEMS

Meaningful Learning: Enhancing Abstract Reasoning in LLMs via Generic Fact Guidance

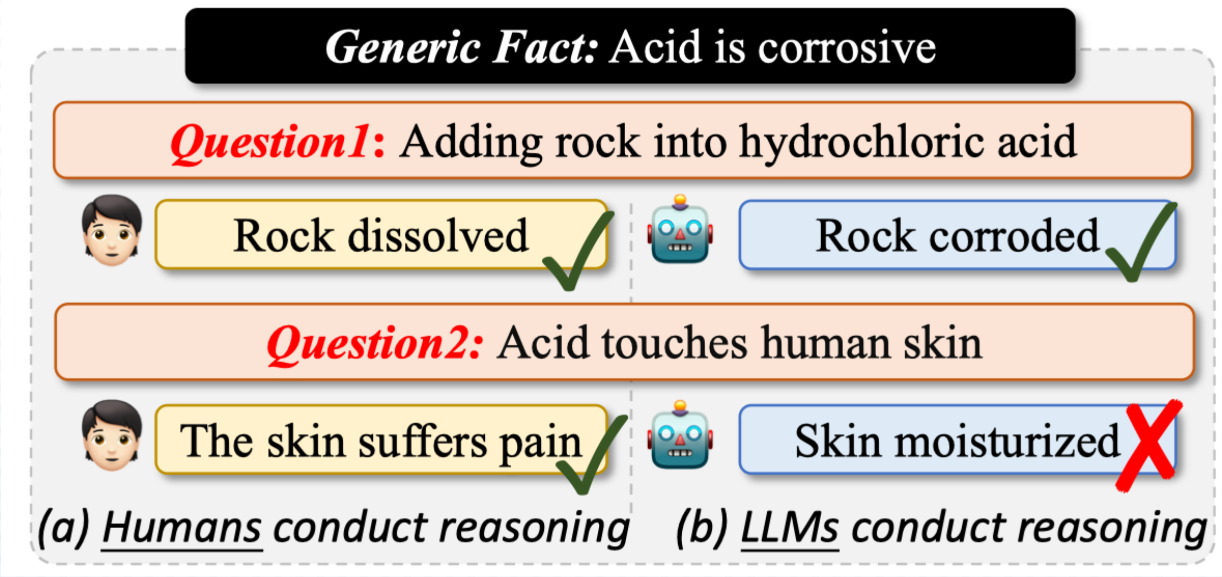
Kai Xiong[1], Xiao Ding[1], Ting Liu[1], Bing Qin[1],
Dongliang Xu[3], Qing Yang[3], Hongtao Liu[3], Yixin Cao[2]

[1]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[2]Fudan University, Shanghai, China
[3]Du Xiaoman (Beijing) Science Technology Co, Ltd., Beijing, China

01 Abstract reasoning

Abstract reasoning

- Definition: abstract reasoning requires models to apply general patterns or high-level abstractions to different scenarios or questions [1].



- When tasked with several simple questions supported by a generic fact, LLMs often struggle to provide consistent and precise answers

[1] Brenden et al., 2017. Building machines that learn and think like people.

02 Pilot study

Metrics: abstract reasoning (AbsAcc)

- Akin to [2], we suppose an LLM $\mathcal{LM}$ has grasped the generic fact r_i if and only if $\mathcal{LM}$ can correctly answer all examples supported by r_i

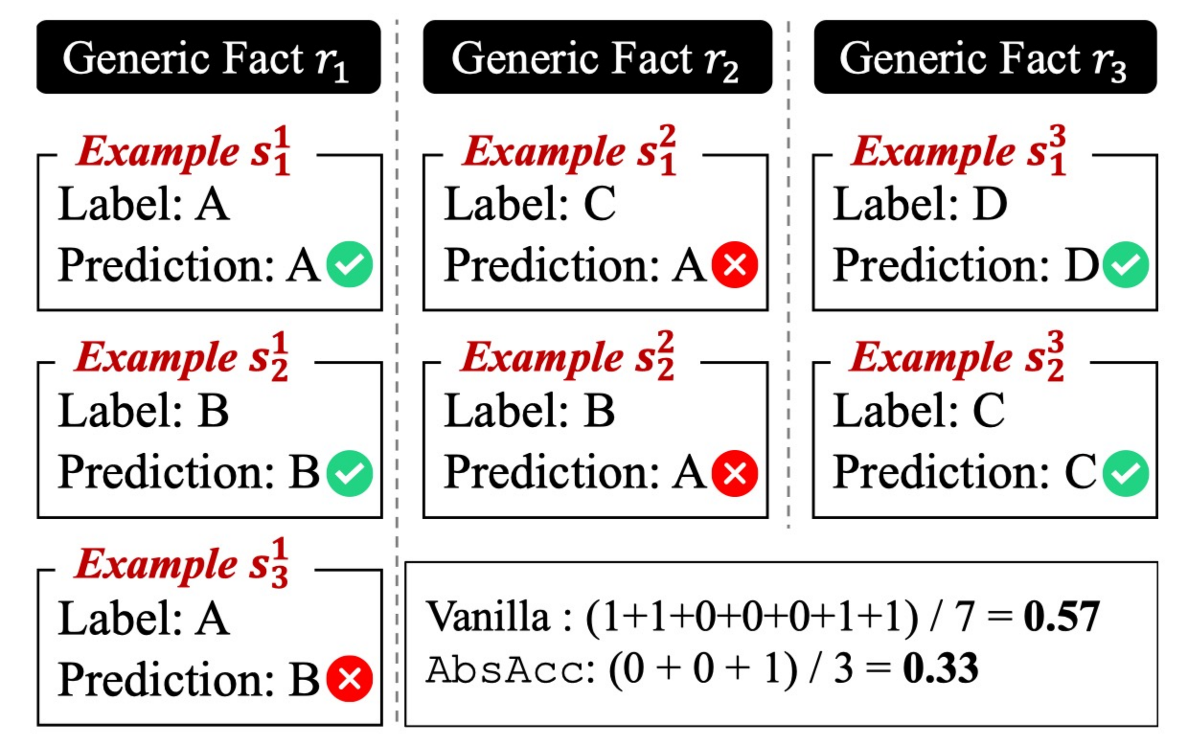


Figure 2: Computation of abstract reasoning metric.

[2] Qiu et al., 2023. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement.

Experiment I: abstract reasoning

- There is a significant disparity (over 17%) between the vanilla accuracy and AbsAcc for all LLMs

Methods	Size	Vanilla Accuracy	AbsAcc
LLaMA-2 [47]	7B	51.67	24.82
	13B	68.45	42.94
Orca-2 [35]	7B	74.23	51.19
	13B	79.83	59.47
GPT-3.5 [7]	>20B	83.25	66.08
Human [14]	–	92.00	89.90

Experiment II: generic fact probing

- Massive reasoning-based post-training (Orca-2) might lose lots of knowledge
- If LLMs know the knowledge, then they can perform well on the corresponding examples

Table 2: Accuracy of generic fact probing.

LLMs	LLaMA-2		Orca-2		GPT-3.5
	7B	13B	7B	13B	>20B
Accu.	62.44	70.32	26.97	12.78	77.30

Table 3: The categorized abstract reasoning accuracy based on whether the generic facts are known.

Category	Metrics	LLaMA-2		Orca-2		GPT-3.5
		7B	13B	7B	13B	>20B
Known	`AbsAcc`	34.26	43.51	64.98	81.51	68.31
Unknown	`AbsAcc`	17.19	31.84	49.00	58.38	58.49

03 Method: meaningful learning

AbsR construction

- Based on GenericsKB [3], we utilize gpt-4-1106-preview to synthesize multiple examples for a single generic fact, which are the applications of the generic fact on different scenarios

Generic Fact: **Unusual conditions affect behavior**

Question: Which of the following scenarios is most likely to influence a person's typical shopping habits?

Options:
A) A regular day with no significant changes in weather or social conditions
B) A holiday season with sales and promotions in stores
C) When the person has not received their paycheck yet
D) A day with consistent weather and social conditions as the previous week

Answer: B)

Explanation: Holiday seasons with promotions and sales present unusual conditions that can significantly alter a person's typical shopping behavior, often encouraging more spending.

Question: In a psychological study, which condition is most likely to yield atypical results in participants?

Options:
A) A study conducted during a stressful event
B) A study conducted in a controlled environment with no distractions
C) A study conducted with the same participants as a previous similar study
D) A study conducted with used questionnaires

Answer: A)

Explanation: A stressful event like a natural disaster creates an unusual condition that can significantly affect participants' behavior and responses, leading to atypical results in a psychological study.

Training

- We employ the constructed AbsR dataset to train LLMs autoregressively, and we model the following two conditional probabilities:

$$\mathcal{LM}(X, Y, \theta) = -\sum_t \log p_\theta(Y_t | X, Y_{<t}),$$

$$\mathcal{LM}(X, r, Y, \theta) = -\sum_t \log q_\theta(Y_t | X, r, Y_{<t}).$$

04 | Experiments

Main results

Size	LLMs	AbsR	AGIEval	RACE	BBH	Com.	MMLU	ARC-e	ARC-c	*Average*
Vanilla Accuracy										
>20B	PaLM-2	75.76	15.83	75.82	48.73	78.05	58.73	89.95	82.37	65.77 ± 0.00
	GPT-3.5	84.00	25.84	83.36	54.15	74.80	65.98	94.71	88.81	71.46 ± 0.00
7B	LLaMA-2	50.00	27.89	38.85	26.00	55.37	41.65	58.73	43.05	42.69 ± 0.00
	Vicuna	75.00	32.56	65.75	31.93	64.77	49.40	74.78	57.63	56.48 ± 0.00
	WizardLM	65.00	22.27	22.11	25.99	43.63	23.26	25.57	22.37	31.28 ± 0.00
	Orca-2	73.50	**38.57**	73.32	34.51	71.58	50.11	79.19	73.90	61.84 ± 0.00
	MeanLearn	**77.00**	38.15	**77.15**	**35.64**	**76.59**	**52.98**	**86.67**	**78.06**	**65.28** ± 0.34
8B	LLaMA-3	82.50	36.33	76.75	36.60	**71.76**	61.86	88.54	75.25	66.20 ± 0.00
	MeanLearn	**84.50**	**38.53**	**77.02**	**37.67**	71.73	**62.35**	**88.89**	**77.97**	**67.12** ± 0.18
13B	LLaMA-2	**73.50**	34.41	59.46	30.61	61.00	**51.87**	71.25	55.25	54.67 ± 0.00
	Vicuna	74.00	33.23	63.54	33.22	63.10	49.34	77.78	59.77	56.75 ± 0.00
	WizardLM	74.00	33.23	63.54	33.22	63.10	49.34	77.78	59.77	56.75 ± 0.00
	Orca-2	66.50	43.75	66.86	39.39	65.92	47.27	82.19	70.85	60.34 ± 0.00
	MeanLearn	67.17	**43.92**	**70.24**	**40.36**	**69.75**	51.68	**88.69**	**82.00**	**64.23** ± 0.25
AbsAcc										
>20B	PaLM-2	64.58	9.59	61.54	27.69	69.34	44.39	85.68	77.10	54.99 ± 0.00
	GPT-3.5	77.08	14.48	66.79	25.45	65.74	52.09	92.16	85.05	59.86 ± 0.00
7B	LLaMA-2	35.42	16.09	14.08	9.56	47.42	25.48	50.27	35.51	29.23 ± 0.00
	Vicuna	58.33	19.90	38.68	14.28	58.07	32.50	65.95	49.53	39.84 ± 0.00
	WizardLM	53.24	12.94	5.57	9.77	32.96	11.55	20.00	17.29	20.42 ± 0.00
	Orca-2	60.42	**26.27**	50.90	15.67	64.02	33.68	72.16	67.76	48.86 ± 0.00
	MeanLearn	**64.58**	25.27	**57.10**	**16.67**	**71.17**	**37.24**	**81.35**	**73.83**	**53.39** ± 0.32
8B	LLaMA-3	72.92	23.78	**55.61**	18.13	66.35	**47.31**	83.51	70.09	54.71 ± 0.00
	MeanLearn	**73.96**	**26.57**	55.09	**19.24**	**66.48**	46.76	**83.78**	**73.36**	**55.66** ± 0.17
13B	LLaMA-2	**61.46**	21.01	32.17	13.74	52.87	**35.72**	62.43	44.86	40.53 ± 0.00
	Vicuna	61.46	19.70	36.50	16.33	55.26	32.21	70.27	52.34	40.37 ± 0.00
	WizardLM	59.38	16.96	28.59	18.61	55.73	29.47	57.84	42.06	38.58 ± 0.00
	Orca-2	50.00	29.50	42.12	20.61	56.41	31.11	75.95	62.15	45.98 ± 0.00
	MeanLearn	45.82	**30.01**	**47.65**	**21.85**	**61.04**	35.21	**85.41**	**77.08**	**50.51** ± 0.23

05 | Conclusion

Conclusion

- We summarize our contributions as follows:
  - We provide a systematic and quantitative analysis of abstract reasoning in current LLMs, and we develop an abstract reasoning dataset with generic-fact-guided explanations.
  - We propose meaningful learning to improve abstract reasoning in LLMs with generic fact.
  - We achieve improvement in general reasoning and abstract reasoning on various OOD reasoning and language understanding benchmarks.

For more details

- If you have any questions, feel free to contact us.

e-mail: kxiong@ir.hit.edu.cn

Wechat: xiongkai_hhh

Telegram: kxionghhh

paper & code

homepage

Thanks!