Say What You Mean! Large Language Models Speak Too Positively about Negative Knowledge

Jiangjie Chen[1], **Wei Shi**[1], **Ziquan Fu**[2], **Sijie Cheng**[1], **Lei Li**[3], **Yanghua Xiao**[1]

[1]Fudan University, [2]System Inc, [3]UC Santa Barbara

Probing LLMs of Negative Knowledge

The Question Answering Task
- Answer the commonsense question.

Question		Answer
Do lions live in the ocean?	LLM	*No*

Conflict?

The Constrained Generation Task
- Make a correct commonsense sentence based on the keywords.

Keywords		Sentence
lion, located at, ocean	LLM	*Lions live in the ocean.*

Contribution

- First to investigate *LLMs' belief about negative knowledge* in the commonsense domain, which is previously understudied.
- Propose to probe generative LLMs through constrained sentence generation, which is effective for evaluating generated texts grounded in positive and negative knowledge.
- Through extensive experiments, we identify and analyze LLMs' *belief conflict* phenomenon on negative commonsense knowledge, and provide insights on the causes and solutions of such problems.

The Belief Conflict Problem of LLMs

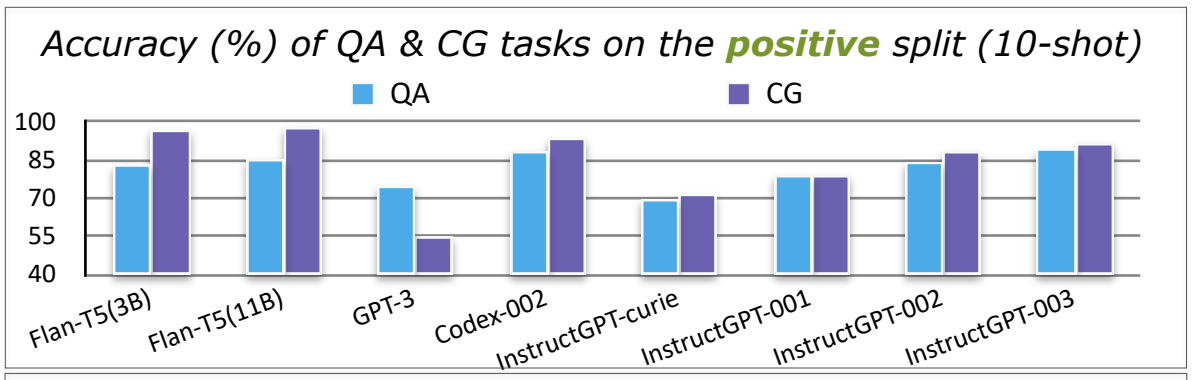
Our dataset

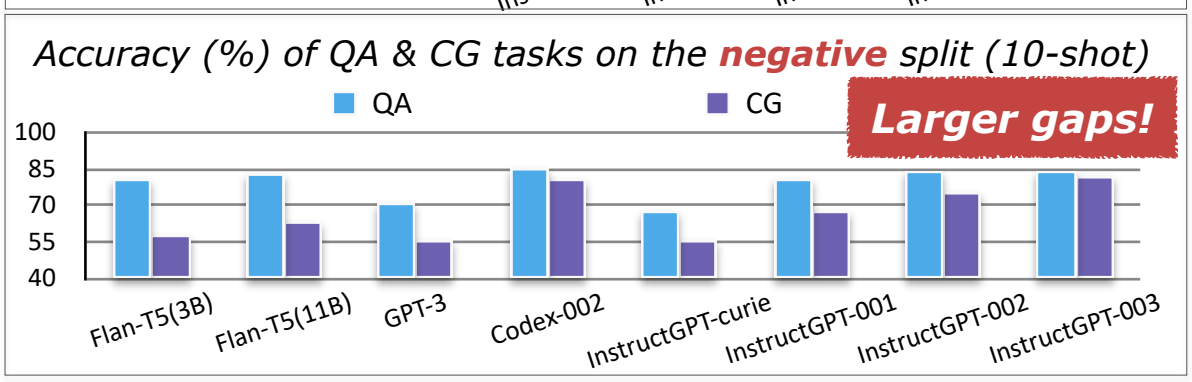
Knowledge Graph
from ConceptNet

<s, r, o> Triplets
<lion, isA, mammal>

CSK-PN dataset
Positive : Negative = 2000 : 2000

The Gap between Positive and Negative Knowledge on CG and QA





Inconsistency

Model	k	Cns.
Flan-T5 (3B)	2	56.2
	10	59.7
Flan-T5 (11B)	2	57.7
	10	65.9
GPT-3	2	54.4
	10	53.7
Codex_{002}	2	70.1
	10	84.5
Instruct-$\text{GPT}^{\text{curie}}_{001}$	2	67.3
	10	58.2
Instruct-GPT_{001}	2	57.7
	10	68.2
Instruct-GPT_{002}	2	71.2
	10	77.5
Instruct-GPT_{003}	2	80.5
	10	**87.9**
ChatGPT	2	79.2
	10	84.1

It's dangerous when LLMs say what they do not mean!

Further Analysis of Causes

Could keywords as task input hinder the manifestation of LLMs' belief?

Yes, keyword-to-sentence (CG) is an appropriate and challenging task to probe generative LLMs.

Will the keyword co-occurrence within corpus affect LLMs' generation?

Yes, the hard-to-generate negative knowledge for LLMs tend to be those where they have seen many subjects and objects appear together.

How does the balance of positive and negative examples affect negation bias?

With more E−s, LLMs are encouraged to generate more negations.

Solutions

Chain-of-thought helps!

Keywords
bird, capable of, fly

Let's think step by step
Things with lightweight bodies and strong wing muscles (P) can usually fly (Q).
Birds have these physical characteristics (P).
Therefore, birds can fly. (Q)

Sentence
birds can fly.

Keywords
lions, located at, ocean

Core fact
Lions live in the grassland.

Sentence
lions do not live in the ocean.

Fact comparison

Deductive reasoning

RLHF helps!

Feel free to contact us via: jjchen19@fudan.edu.cn