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

Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng,
Lei Li, Yanghua Xiao

復旦大學 FUDAN UNIVERSITY
Brain Technologies Inc.
UC SANTA BARBARA

Commonsense knowledge and LLMs: Both positive and negative

lions live in grasslands

lions don't live in the ocean.

Positive*:

Everything that exists is positive.

Negative*:

What is not true.

What cannot be done.

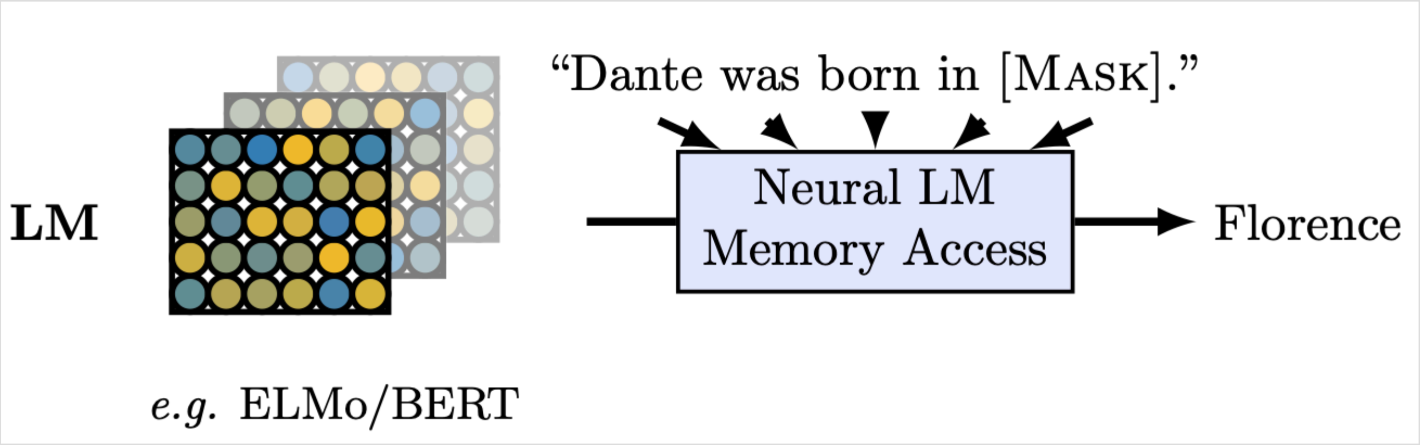
What does not exist.

…

*: (Molnar, 2000; Barker and Jago, 2012)

Do LLMs acquire implicit negative commonsense knowledge?

Mask-infilling task, e.g., LAMA

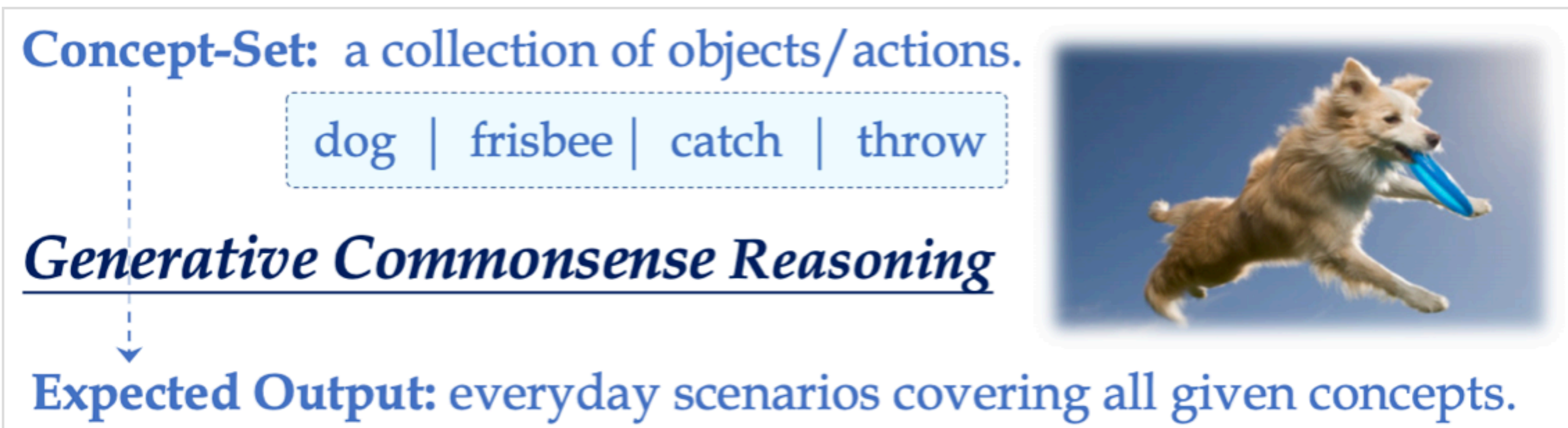


- Not natural for unidirectional LLMs
- Suffers from the open-world problem in evaluation

Language Models as Knowledge Bases? Petroni et al. 2019

Can LLMs generate sentences grounded in such knowledge?

Knowledge-grounded text generation, e.g., CommonGen

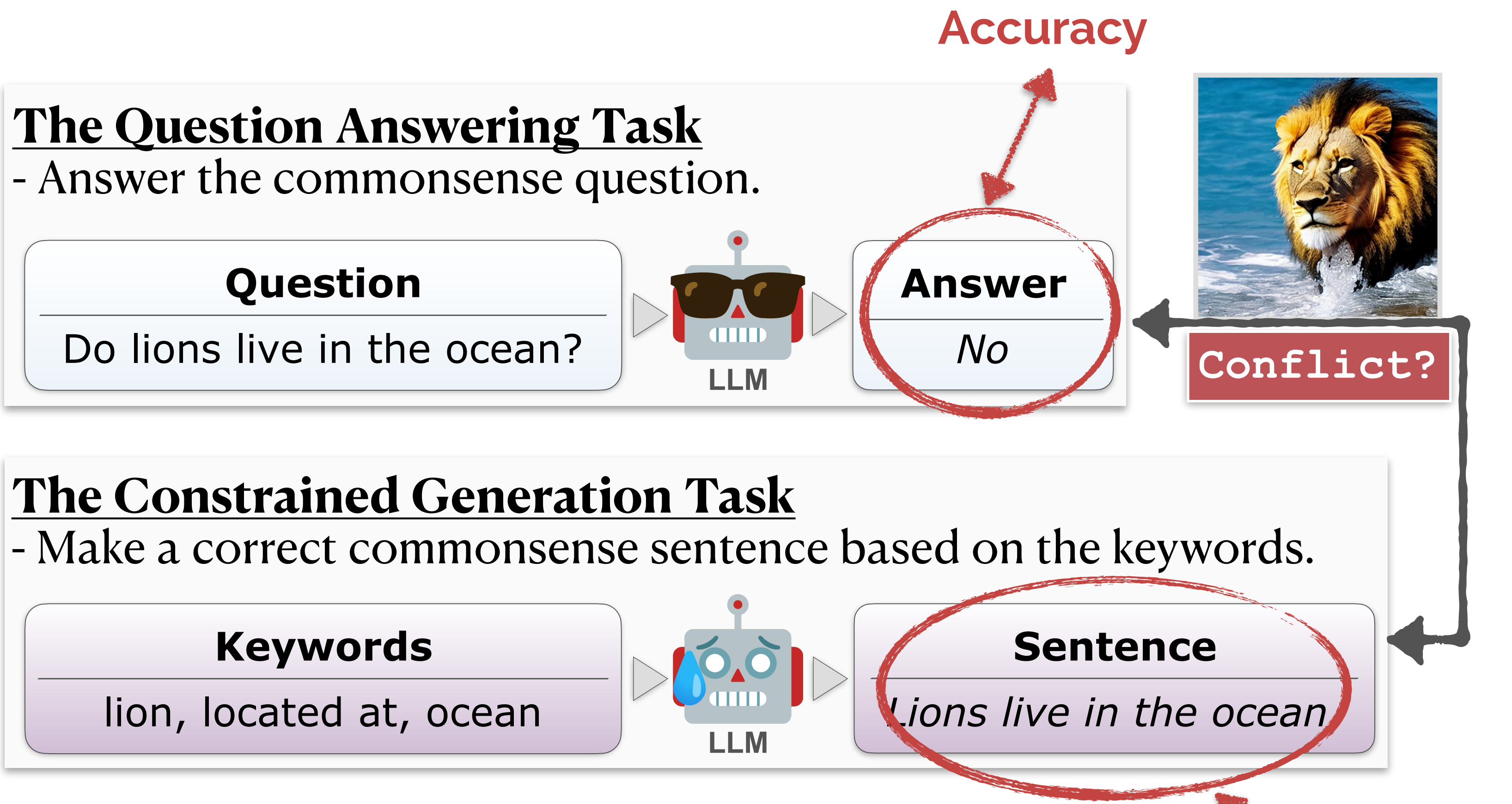


- Do not investigate generating negative knowledge.

CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. Lin et al. 2020

How to probe a generative LLM with negative knowledge?

Two Tasks for Probing Negative Knowledge in LLMs

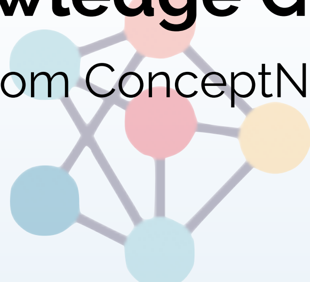


Probe LLMs with **in-context learning.**

Composition of Probing Data

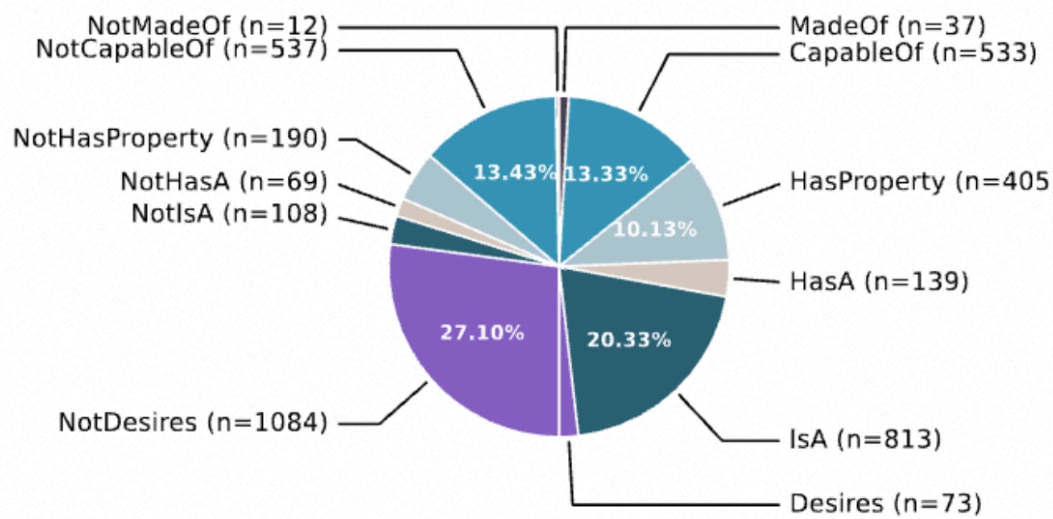
Knowledge Graph

from ConceptNet



<s, r, o> Triplets

<lion, isA, mammal>

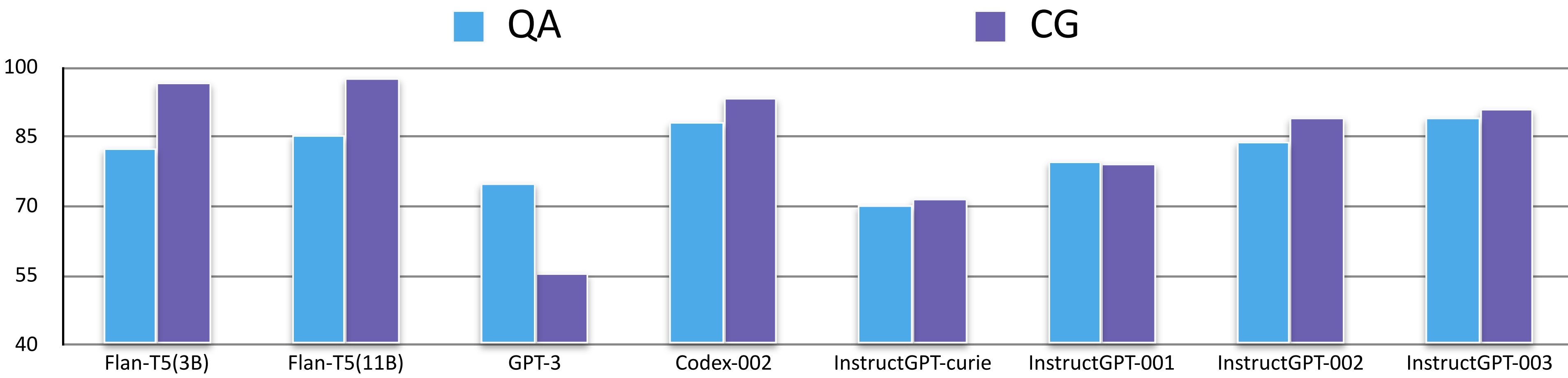


CSK-PN dataset

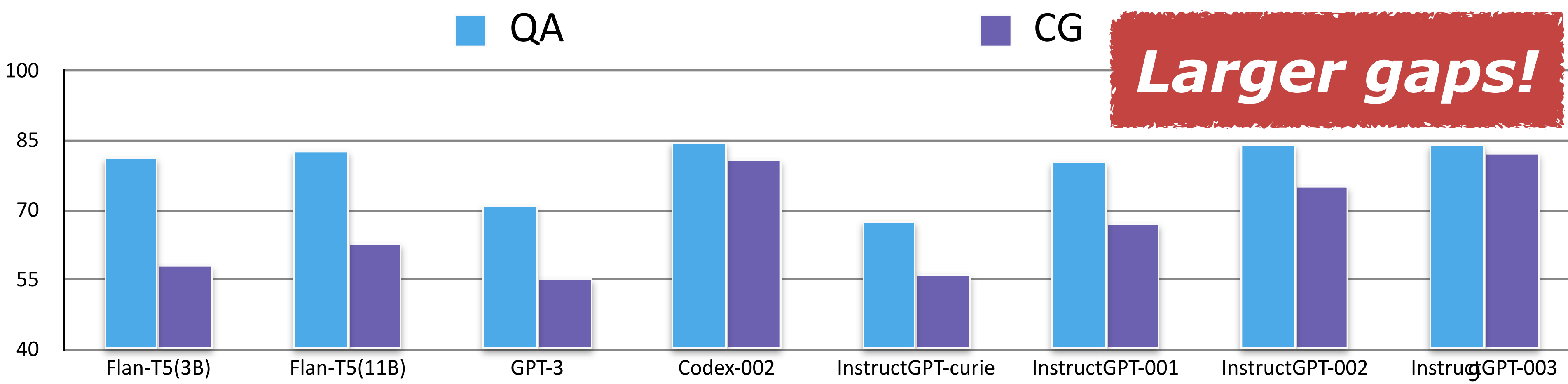
Positive : Negative = 2000 : 2000

Do LLMs have negative knowledge?

The Gap between Positive and Negative Knowledge on CG and QA

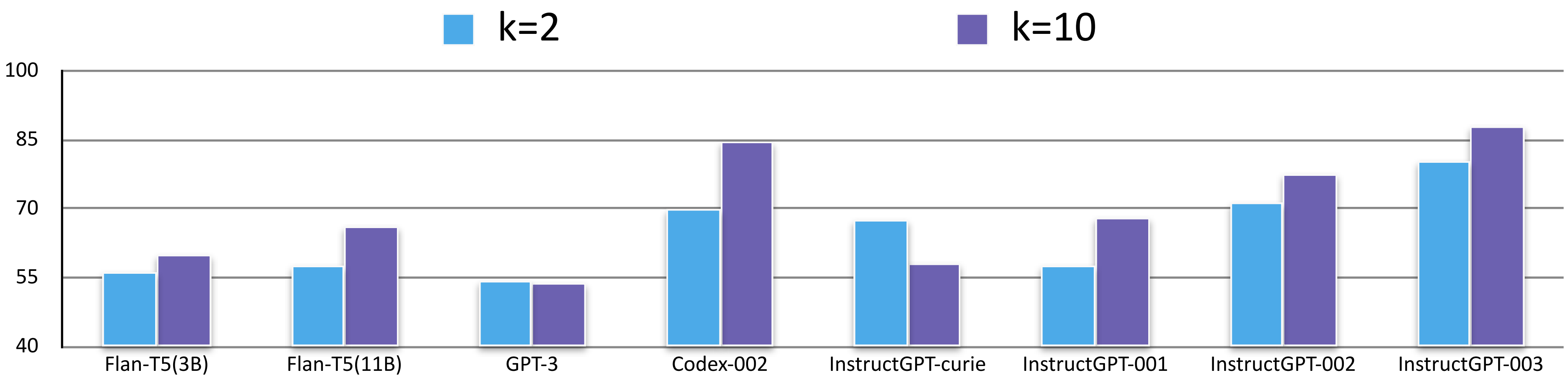


Accuracy (%) of QA & CG tasks on the ***positive*** *split (10-shot)*



Accuracy (%) of QA & CG tasks on the ***negative*** *split (10-shot)*

Consistency between CG and QA



Consistency between QA & CG tasks (k-shot)

Question

Do lions live in the ocean?

Answer

No

Conflict!

Keywords

lion, located at, ocean

Sentence

Lions live in the ocean.

The “Belief Conflict”

- LLMs frequently **fail the CG task** by generating invalid sentences grounded in negative commonsense knowledge.
- But LLMs can **correctly answer the QA questions**, demonstrating they know the negative knowledge.
- **It’s dangerous when LLMs say what they do not mean.**

What are the Causes of Belief Conflict?

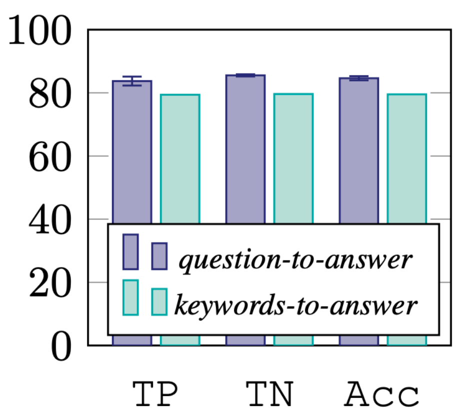
Could keywords as task input hinder the manifestation of LLMs' belief?

```
Answer the question by writing a short
sentence that contains correct common
sense knowledge.

Question: do lions live in the ocean?
Sentence: lions don't live in the ocean.
```

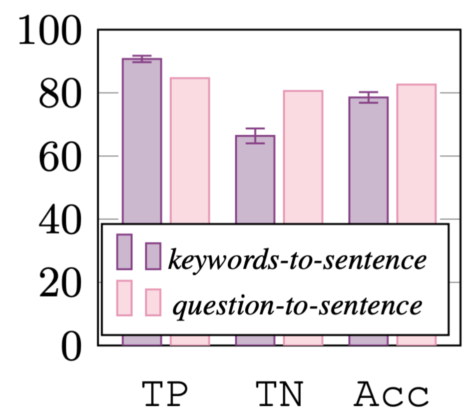
```
Can these keywords form a truthful
common sense fact? Answer with yes or
no.

Keywords: lion, located at, ocean
Answer: no
```



(a) Results (%) on QA.

Switch input in CG & QA tasks:



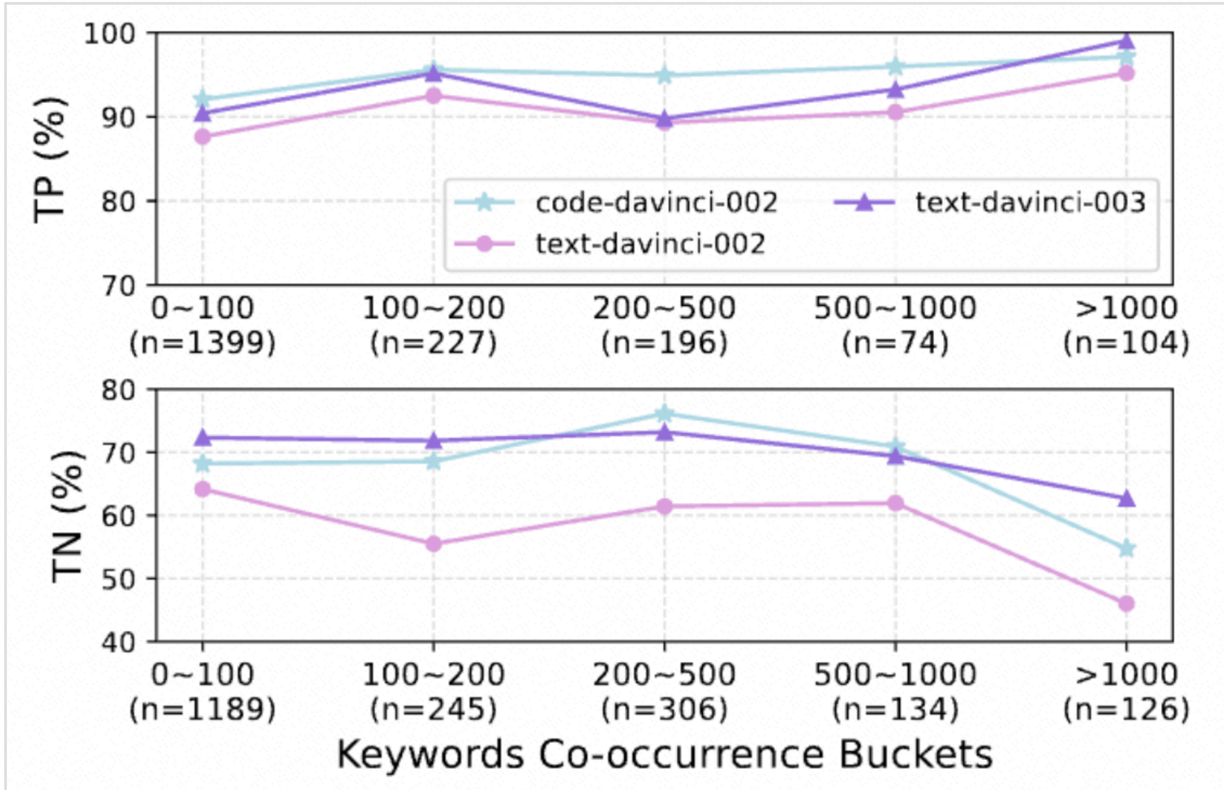
(b) Results (%) on CG.

1. keyword-to-sentence (CG) is an appropriate and challenging task to probe generative LLMs.
2. Keyword inputs for negative knowledge do not have a statistical shortcut from pre-training.

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

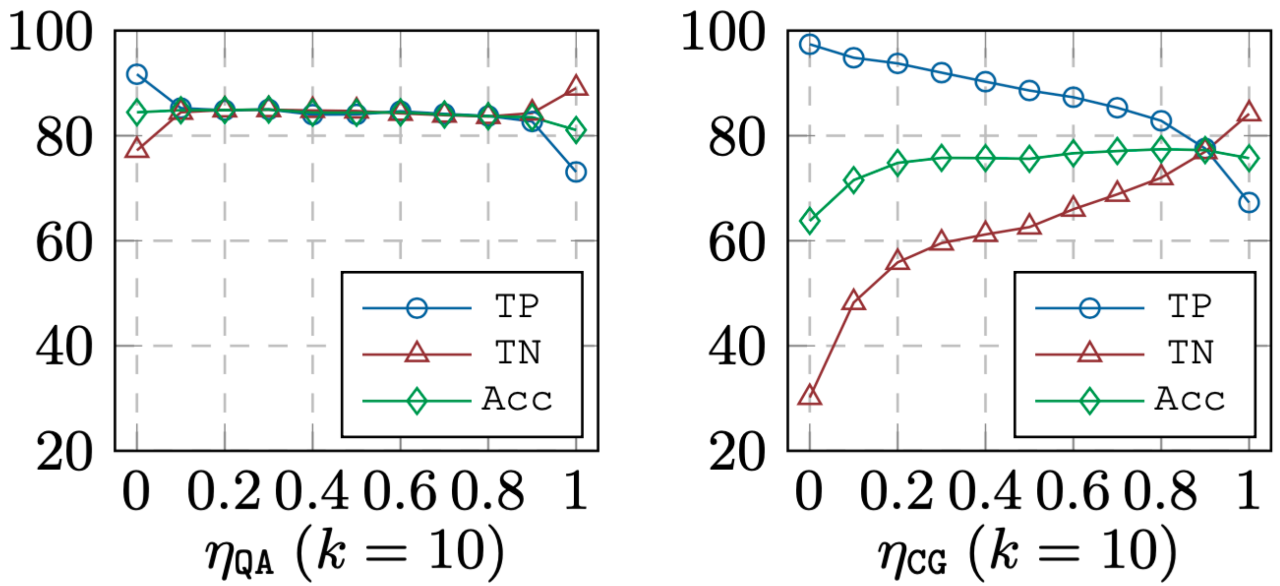
Co-occurrence:

$$\frac{\sum_{i,j} \text{cooccur}(w_i, w_j)}{l_s l_o}$$



1. 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?



1. With more E–s, LLMs are encouraged to generate more negations.
2. The belief conflict can be overcome by increasing negated texts in the training data or in-context examples. (Not always feasible.)

How to Alleviate the Belief Conflict?

Chain-of-Thought Helps 🤩: Deductive Reasoning

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.

If P then Q. Not Q. Therefore, Not P.

If P then Q. P. Therefore, Q.

Model	CoT	$k = 2$ (1:1)			$k = 10$ (1:1)		
		TP	TN	Acc	TP	TN	Acc
Codex$_{002}$	None	**96.6**	38.0	67.3	**93.2**	68.8	81.0
	Deduction	86.9	**56.6**	71.7	83.5	73.0	78.3
Instruct-GPT$_{002}$	None	**92.9**	51.4	72.1	**88.9**	61.4	75.1
	Deduction	87.0	**57.3**	72.1	84.3	**70.7**	**77.5**

Chain-of-Thought Helps 🤩: Fact Comparison

Keywords

lions, located at, ocean

Core fact

Lions live in the grassland.

Sentence

lions do not live in the ocean.

Model	CoT	$k = 2$ (1:1)			$k = 10$ (1:1)		
		TP	TN	Acc	TP	TN	Acc
Codex$_{002}$	None	**96.6**	38.0	67.3	**93.2**	68.8	81.0
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Fact	92.9	53.7	**73.3**	86.8	**76.6**	**81.7**
Instruct-GPT$_{002}$	None	**92.9**	51.4	72.1	**88.9**	61.4	75.1
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Fact	89.1	55.5	**72.2**	85.5	69.2	77.4

1. Even though LLMs picked up implicit bias during pre-training, it can be overcome by making the reasoning chain explicit.
2. LLM holding concerns of exceptions? Yes, but the conclusion still stands.

RLHF (Somehow) also Helps 🤔

Model	k	Perf. on QA			Perf. on CG			Cns.
		TP	TN	Acc	TP	TN	Acc	
Instruct-GPT$_{002}$	2	81.7	**86.1**	83.9	92.9	48.7	72.1	71.2
	10	84.1	<u>84.7</u>	84.4	88.9	61.4	75.1	77.5
Instruct-GPT$_{003}$	2	87.9	81.3	84.6	95.1	58.1	76.6	80.5
	10	<u>89.0</u>	79.5	84.2	91.1	73.6	<u>82.3</u>	**87.9**
ChatGPT	2	82.9	82.0	82.4	89.8	69.8	79.8	79.2
	10	81.5	85.7	83.6	90.4	<u>78.4</u>	**84.4**	84.1

consis tency

1. Models with RLHF (InstructGPT-003, ChatGPT) are better and more consistent at QA and CG.
2. Negative knowledge and rebuttal statements are frequently used in human feedback to steer the model?
3. Does RLHF lead to cheating?

Say What You Mean!

Large Language Models Speak Too Positively about Negative Commonsense Knowledge

Jiangjie Chen, Wei Shi, Ziquan Fu,
Sijie Cheng, Lei Li, Yanghua Xiao

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

More details in the paper!