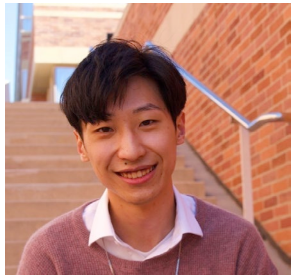
Harnessing Black-Box Control to Boost Commonsense in LMs' Generation

Yufei Tian, Felix Zhang, Nanyun (Violet) Peng



Plus lab

Motivations -- Why?

Challenge 1 - LLMs are unreliable and fail to generate commonsensical outputs at times

(a) Concepts	***wear, sunglasses, at night***
□ **GPT-2**	*A young woman wearing a long dress and ~~sun~~glasses at night.*
□ **Alpaca**	*We wore our sunglasses at night and enjoyed the stars.*

(b) Concepts	***food, customer, watch, employee, prepare***
□ **GPT-2**	*Two employees watch as customers prepare food in the store.*
□ **GPT-3 Davinci-003**	*The employee watched as the customer prepared their food.*

Table 1. Examples of generative commonsense reasoning. We highlight the insensible phrases in orange.

Challenge 2 - It is computationally difficult for many parties to finetune PTLMs with billions of parameters.

Our solution – How?

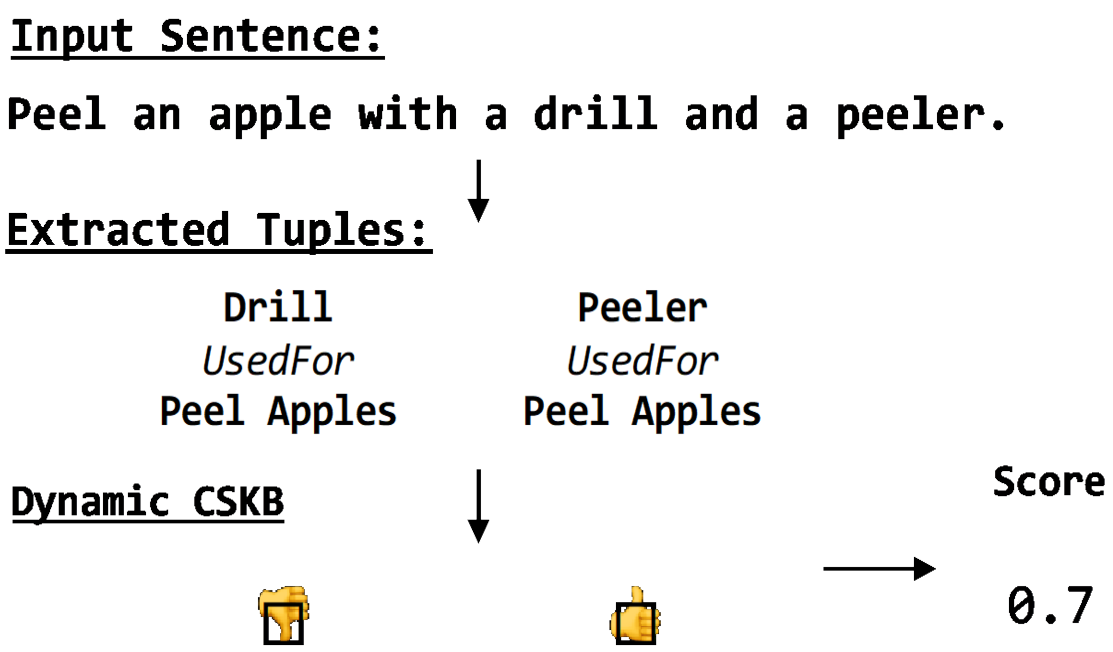
A computational efficient way *to improve the commonsense* of pre-trained language models *in a plug-and-play manner.*

1. We build a reference-free scorer that evaluates how CS a sentence is.
2. (Based on the recent development of controllable generation…)

We train a small auxiliary model to control a frozen PTLM by training on its *self-generated* samples.

Build Commonsense Scorer

- Step 1: ***extract*** tuples from a sentence
- Step 2: ***assign*** each tuple with a score by ***grounding*** them to a dynamic commonsense knowledge base.
- Step 3: The sentence-level score is then obtained by *aggregating tuple-level* scores.



Commonsense-Guided Generation

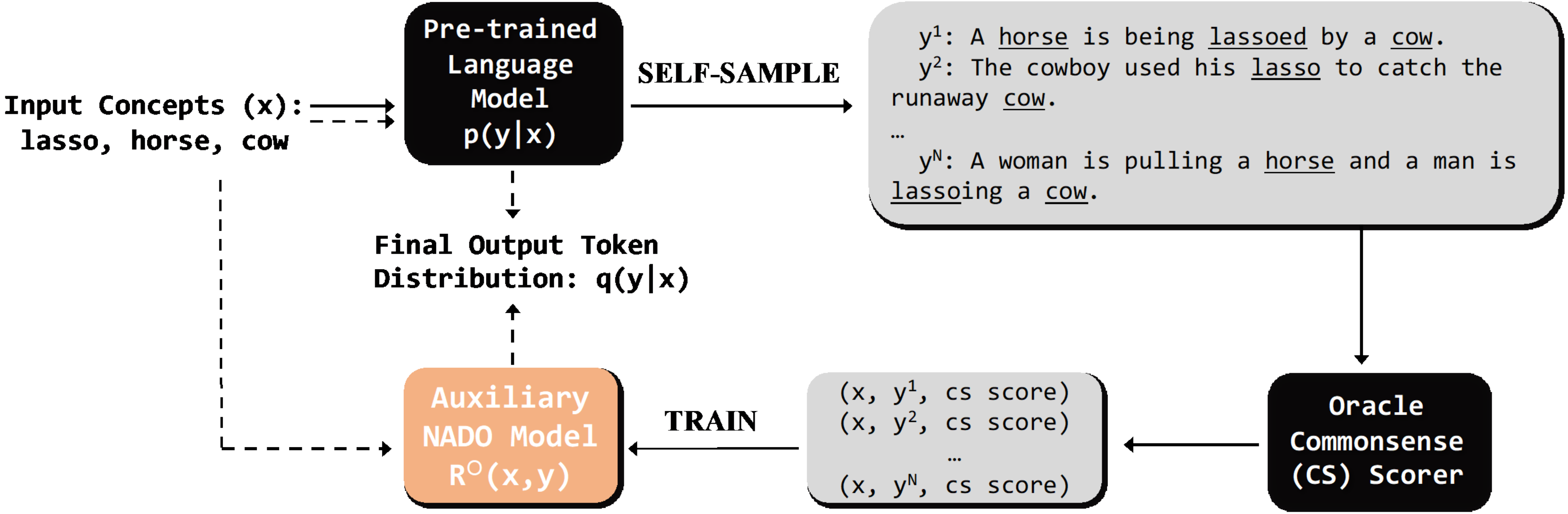


Figure 3. The process to steer a frozen PTLM with an additional neural model (NADO) and oracle commonsense scorer. The solid lines indicate the training process; the dashed lines indicate inference.

Results

1. Our approach is ***effective*** on across three different model architectures, with and without instruction tuning.

2. Few-Shot settings surpass finetuning – ***quality*** of dataset is more important than *size*

3. Human annotators find *ChatGPT more commensensical, but less exciting*

Winning System	BOOST CS	Same	ChatGPT
CS	30%	17%	53%
Overall	47%	25%	28%

Table 2. Pair-wise comparison with ChatGPT.

Comparison with ChatGPT

ChatGPT is too correct → human have less fun.

Constraint: ***wear, sunglasses, at night***

- *ChatGPT: "It is not advisable to wear sunglasses at night as it can impede your vision and increase the risk of accidents."*
- *Boost: Someone wears sunglasses at night to avoid the bright lights of the approaching car.*

Thank you!