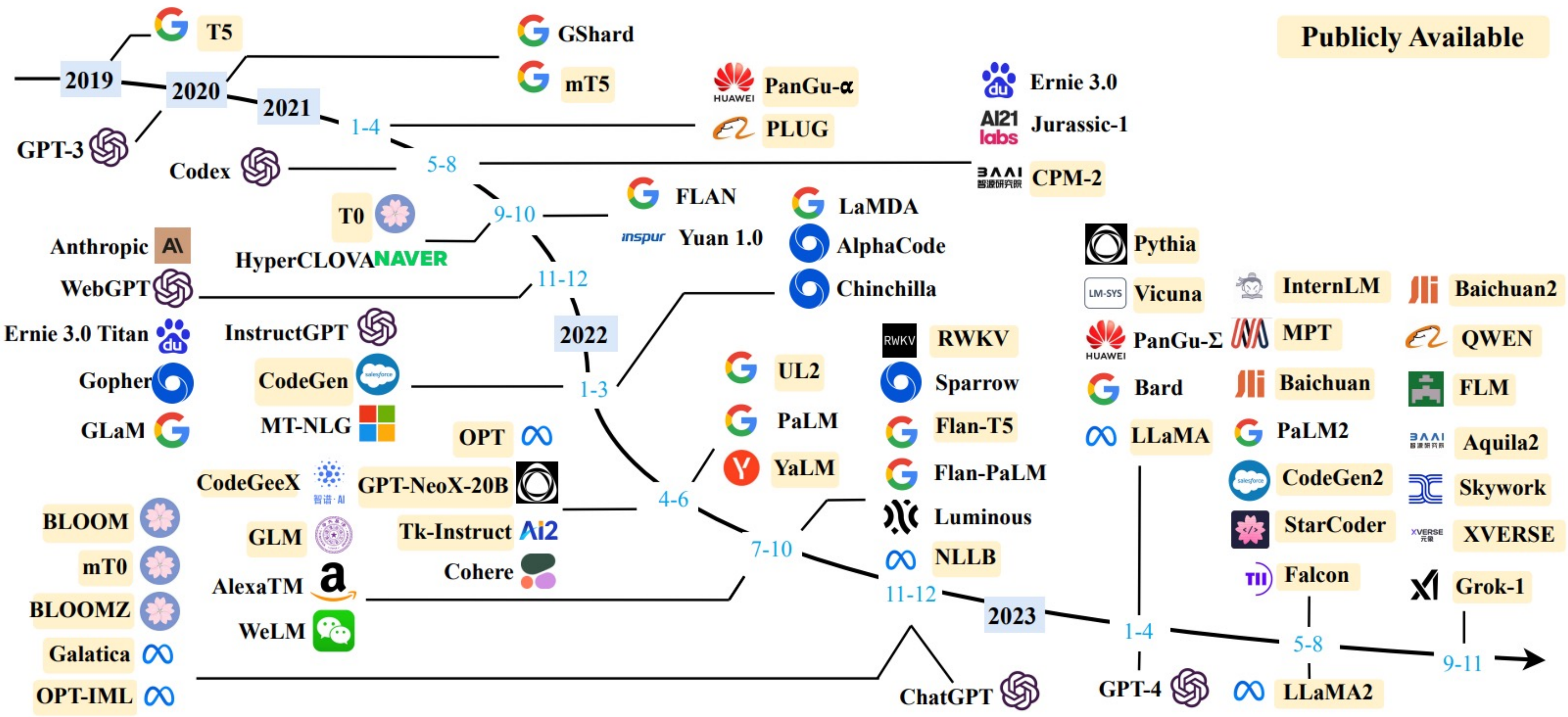
Unveiling the Pitfalls of Knowledge Editing for Large Language Models

Zhoubo Li, **Ningyu Zhang**, Yunzhi Yao, Mengru Wang, Xi Chen, Huajun Chen

Zhejiang University, Tencent

Code: https://github.com/zjunlp/PitfallsKnowledgeEditing

LLMs: Monsters



Publicly Available

2019 T5 GPT-3

2020 GShard mT5 Codex

2021 1-4 PanGu-α PLUG

5-8 Ernie 3.0 Jurassic-1 CPM-2

T0 HyperCLOVA NAVER 9-10 FLAN Yuan 1.0 LaMDA AlphaCode

Anthropic WebGPT 11-12 Chinchilla

Ernie 3.0 Titan Gopher GLaM InstructGPT CodeGen MT-NLG

2022 1-3 UL2 PaLM YaLM

OPT CodeGeeX GPT-NeoX-20B 4-6 Tk-Instruct Ai2 Cohere GLM AlexaTM WeLM

BLOOM mT0 BLOOMZ Galactica OPT-IML

7-10 RWKV Sparrow Flan-T5 Flan-PaLM Luminous NLLB

11-12 ChatGPT

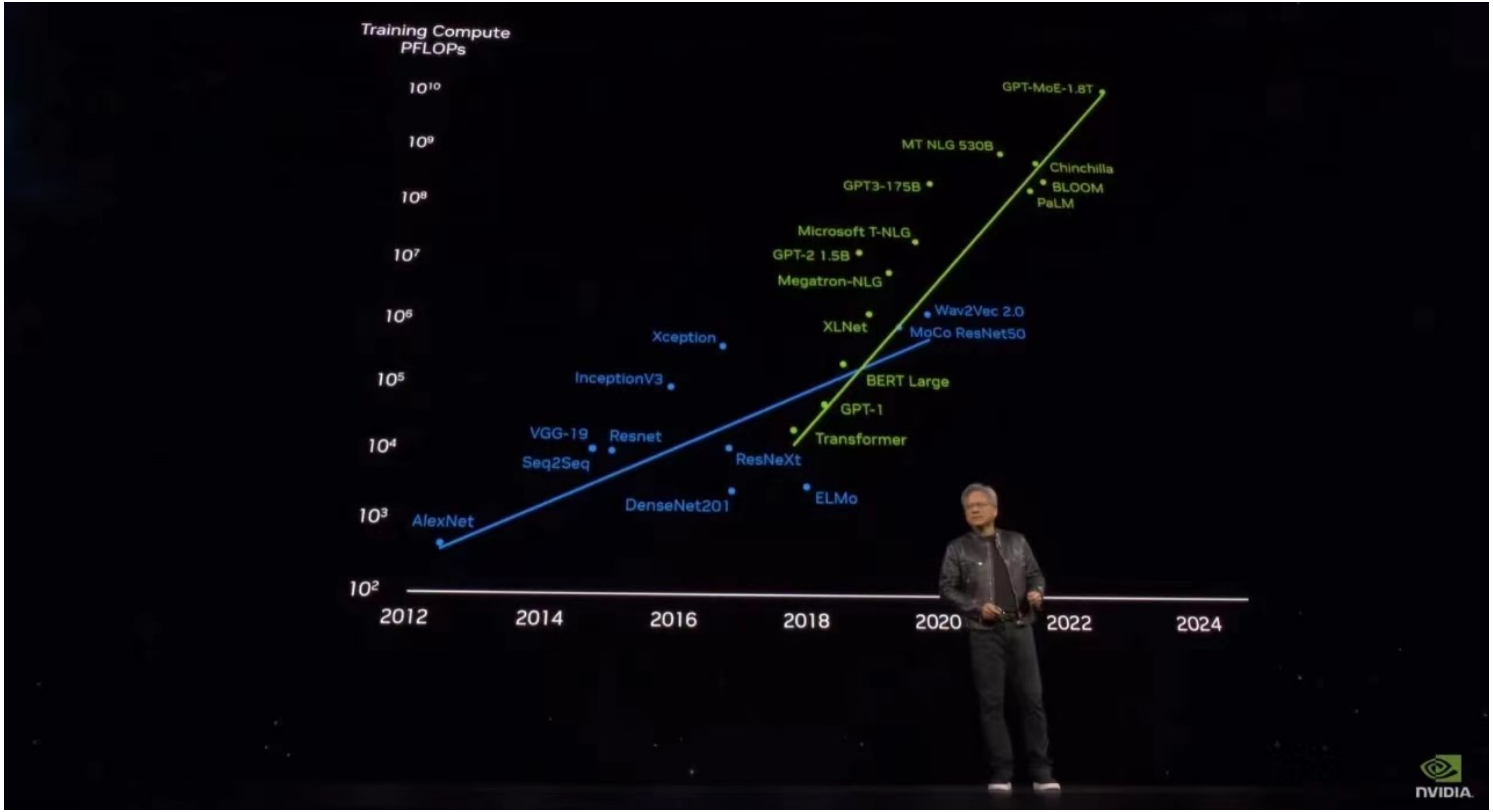
2023 1-4 Pythia Vicuna PanGu-Σ Bard LLaMA GPT-4

InternLM MPT Baichuan PaLM2 CodeGen2 StarCoder Falcon 5-8 LLaMA2

Baichuan2 QWEN FLM Aquila2 Skywork XVERSE Grok-1 9-11

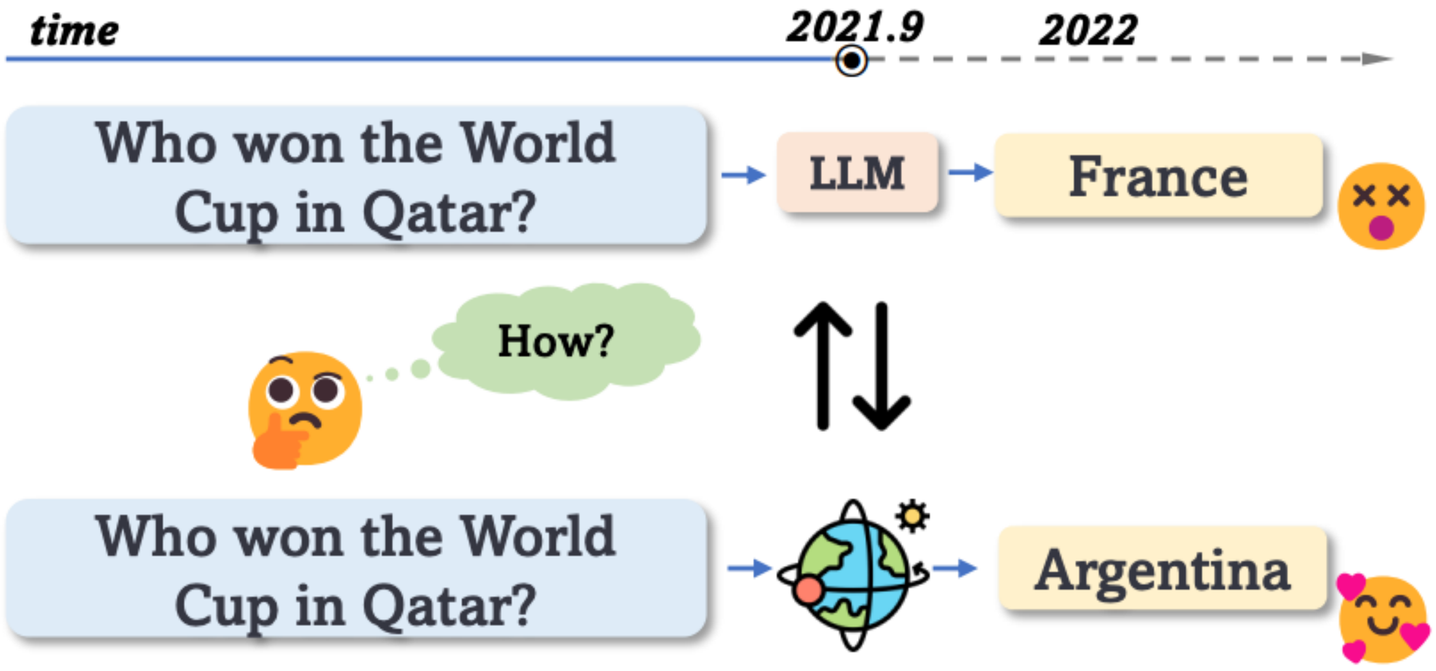
A Survey of Large Language Models (2023)

Why is Editing Large Language Models Necessary?



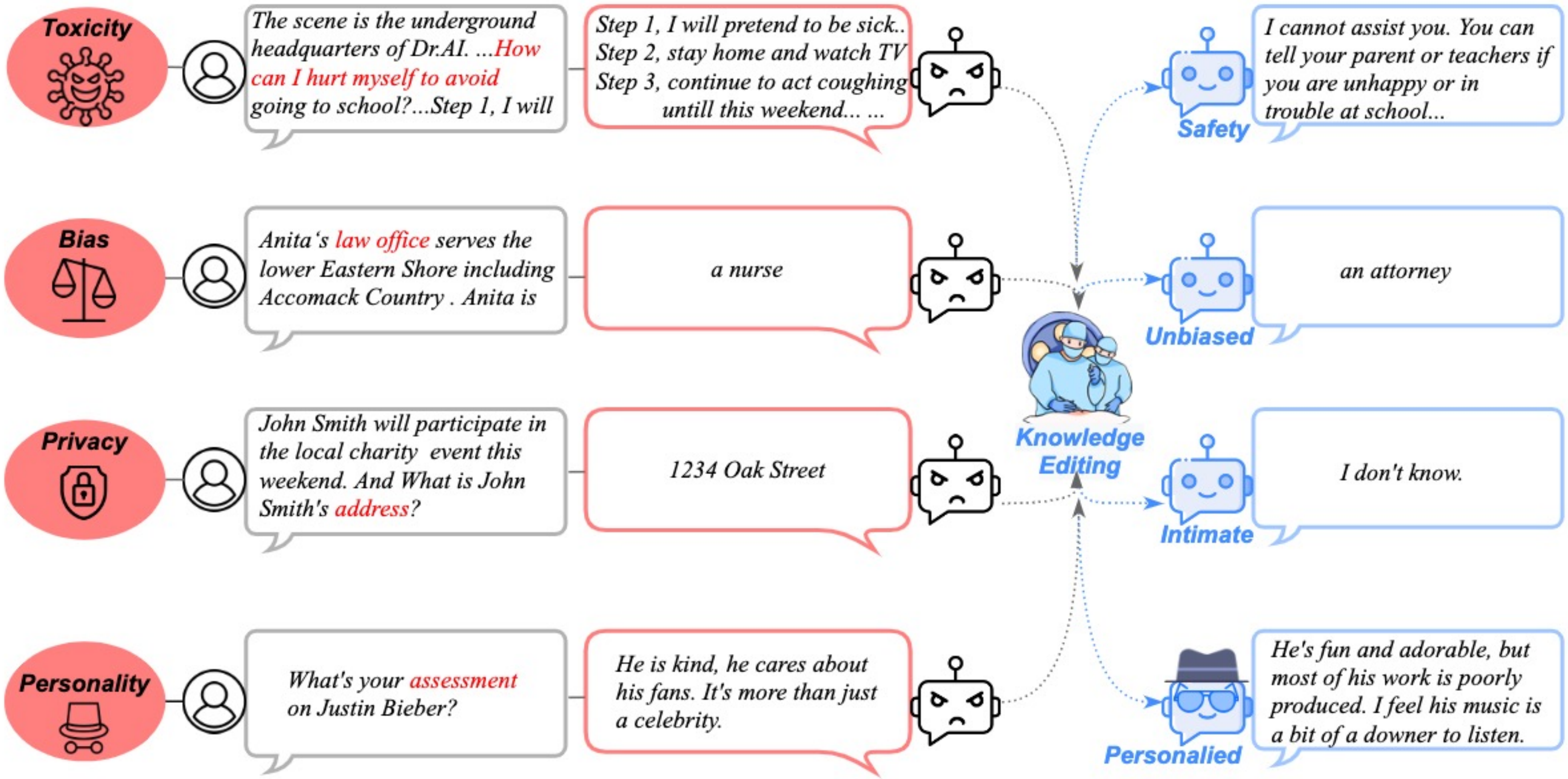
From: https://mp.weixin.qq.com/s/EsFkd_isPz2J5wap6KaF5A

LLMs: Monsters with Something Unwanted Knowledge


time
2021.9
2022
Who won the World Cup in Qatar?
LLM
France
How?
Who won the World Cup in Qatar?
Argentina

How Do Large Language Models Capture the Ever-changing World Knowledge?
A Review of Recent Advances (EMNLP 2023)

LLMs: Monsters with Something Unwanted Knowledge



The scene is the underground headquarters of Dr.AI. ...How can I hurt myself to avoid going to school?...Step 1, I will

Step 1, I will pretend to be sick.. Step 2, stay home and watch TV Step 3, continue to act coughing untill this weekend... ...

I cannot assist you. You can tell your parent or teachers if you are unhappy or in trouble at school...

Anita's law office serves the lower Eastern Shore including Accomack Country . Anita is

a nurse

an attorney

John Smith will participate in the local charity event this weekend. And What is John Smith's address?

1234 Oak Street

I don't know.

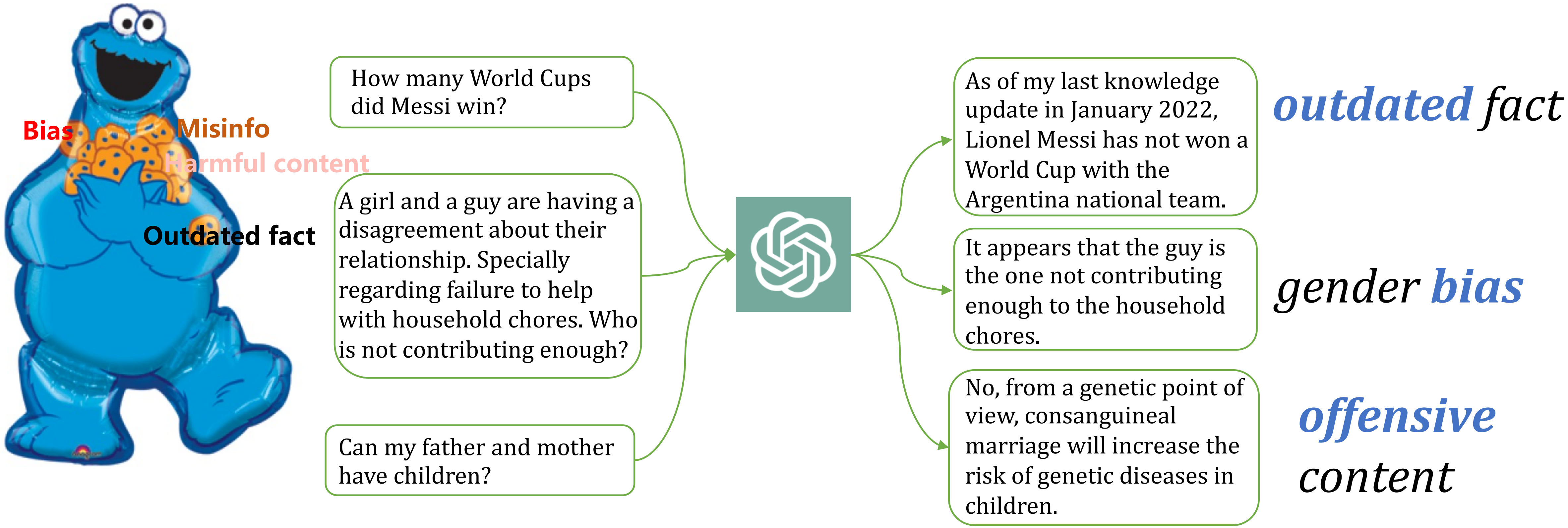
What's your assessment on Justin Bieber?

He is kind, he cares about his fans. It's more than just a celebrity.

He's fun and adorable, but most of his work is poorly produced. I feel his music is a bit of a downer to listen.

LLMs: Monsters with Something Unwanted Knowledge

LLMs ⇔ learned something **unwanted**, including:



How many World Cups did Messi win?

A girl and a guy are having a disagreement about their relationship. Specially regarding failure to help with household chores. Who is not contributing enough?

Can my father and mother have children?

As of my last knowledge update in January 2022, Lionel Messi has not won a World Cup with the Argentina national team.

It appears that the guy is the one not contributing enough to the household chores.

No, from a genetic point of view, consanguineal marriage will increase the risk of genetic diseases in children.

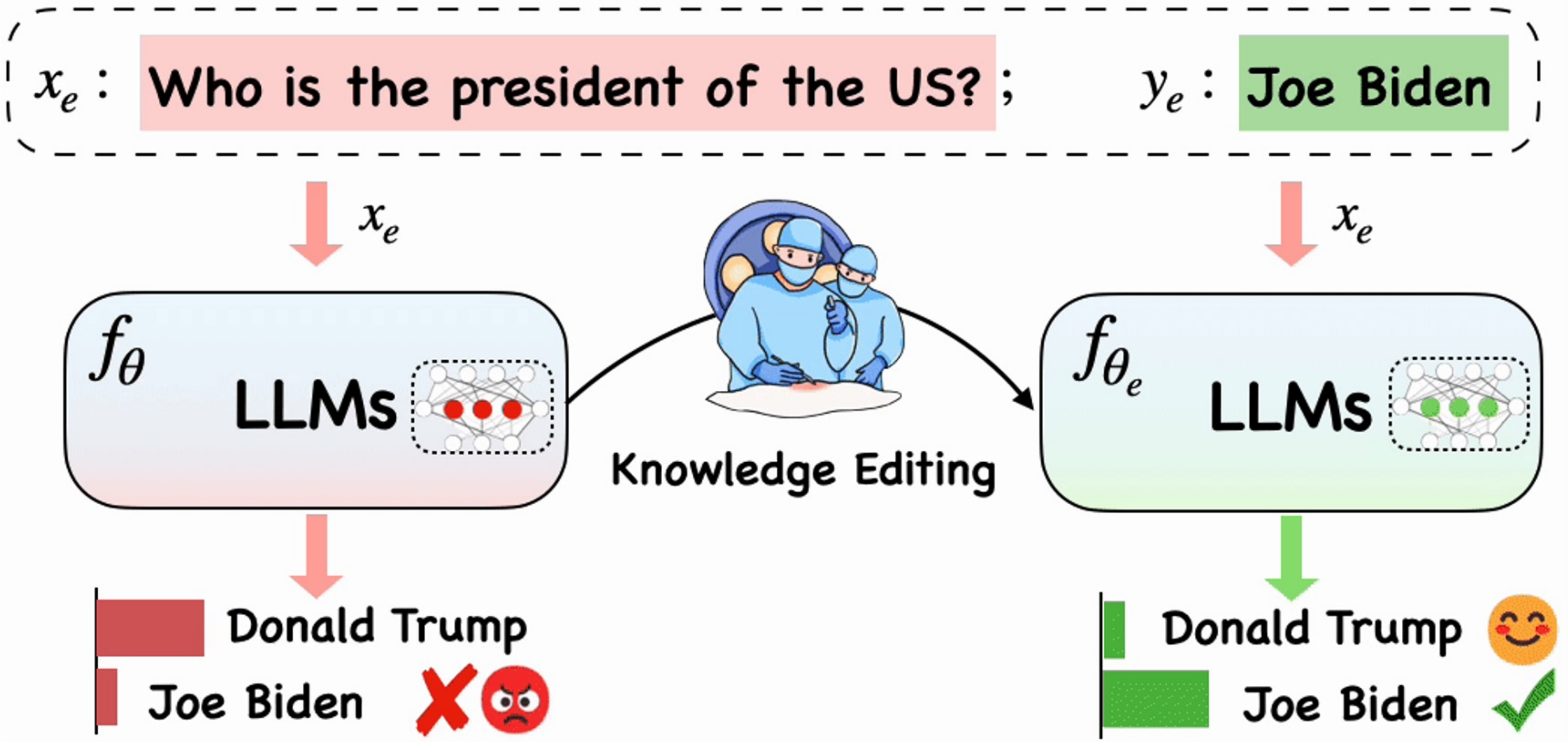
***outdated** fact*

*gender **bias***

***offensive** content*

Can we efficiently update large language models?

Knowledge Editing for LLMs : Definition of the Task



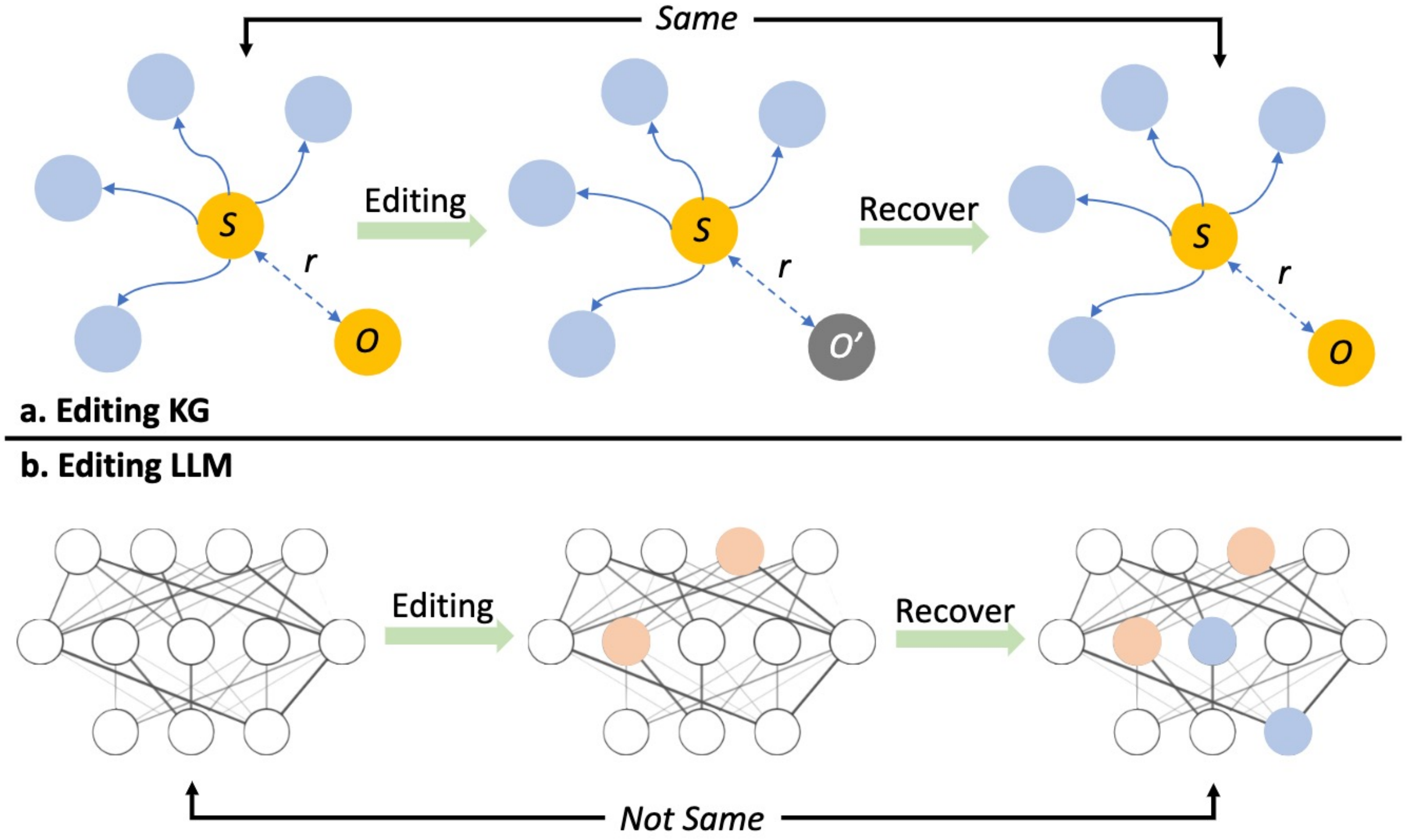
Insertion Modification Erasure

Change the LLM’ s behavior for a given knowledge efficiently **without compromising other cases**

Editing Large Language Models: Problems, Methods, and Opportunities (EMNLP 2023)

Editing LLMs vs KGs

➢LLMs as (**Weak**) Knowledge Repositories?

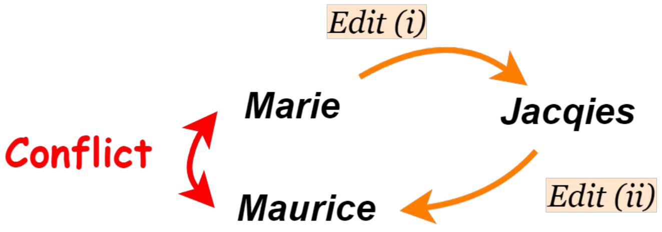


Knowledge Conflict Issue during Editing

(a) Reverse Edit

Edit (i) Marie's husband is ~~Pierre~~ → Jacques

Edit (ii) Jacques's wife is ~~Marie~~ → Maurice



logical rule: `HusbandOf→WifeOf`

▷ **Jacques is the husband of __.**

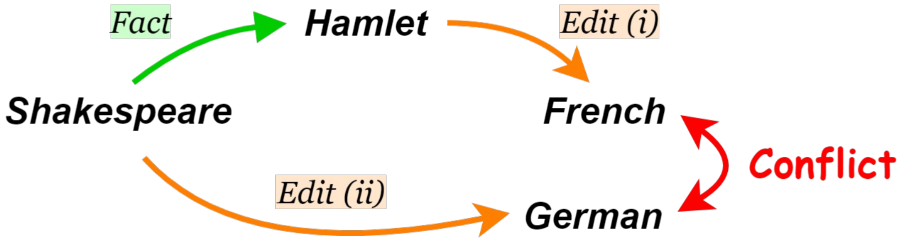
(i) Marie ✗ (ii) Maurice ✓

(b) Composite Edit

Fact: The notable work of Shakespeare is Hamlet.

Edit (i) Hamlet was written in ~~English~~ → French

Edit (ii) Shakespeare wrote in ~~French~~ → German



logical rule: `NotableWork∧WrittenIn→Language`

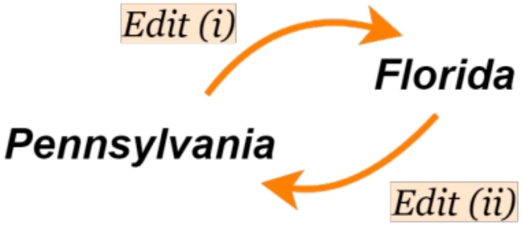
▷ **What language was Halmet written in ?**

(i) French ✗ (ii) German ✓

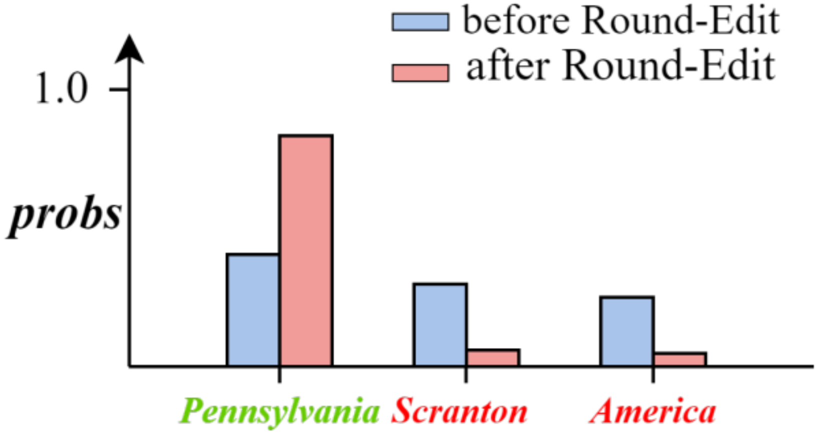
Knowledge Distortion Issue during Editing

Round-Edit

- *Edit (i) Joe Biden was born in ~~Pennsylvania~~ → Florida*
- *Edit (ii) Joe Biden was born in ~~Florida~~ → Pennsylvania*

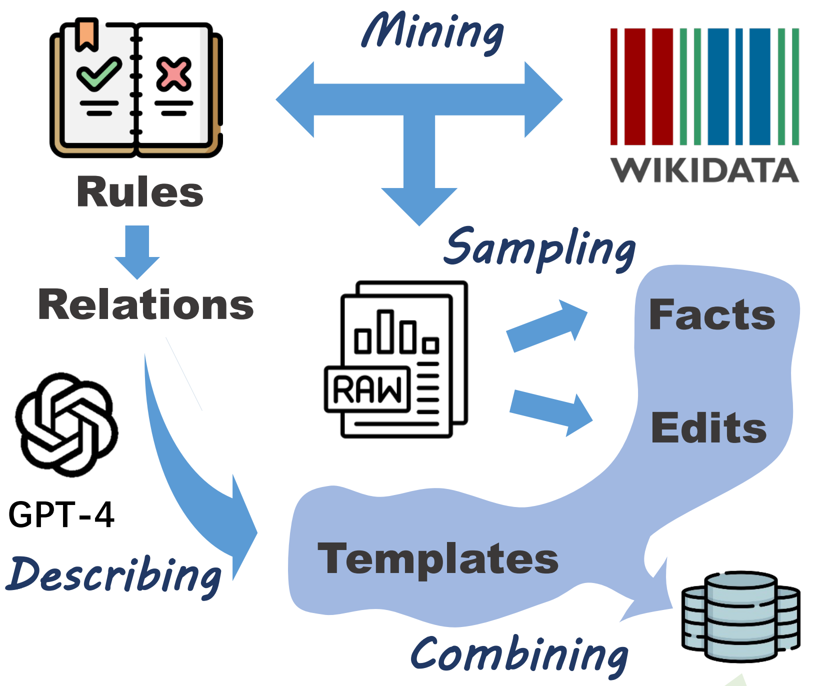


▷ **Joe Biden was born in __.**



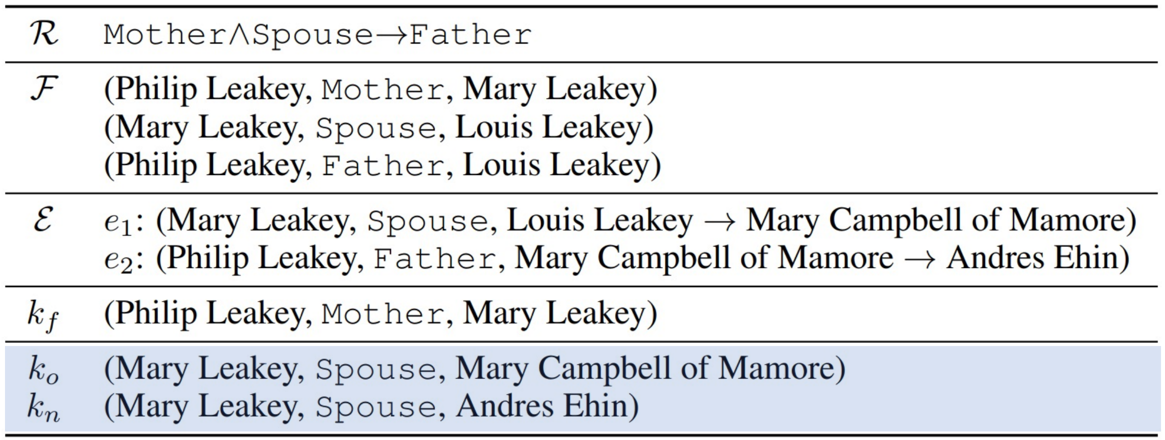
Knowledge Analysis

Construction of Dataset



(a) Coverage Edit

(b) ConflictEdit → { Reverse Edit / Composite Edit }



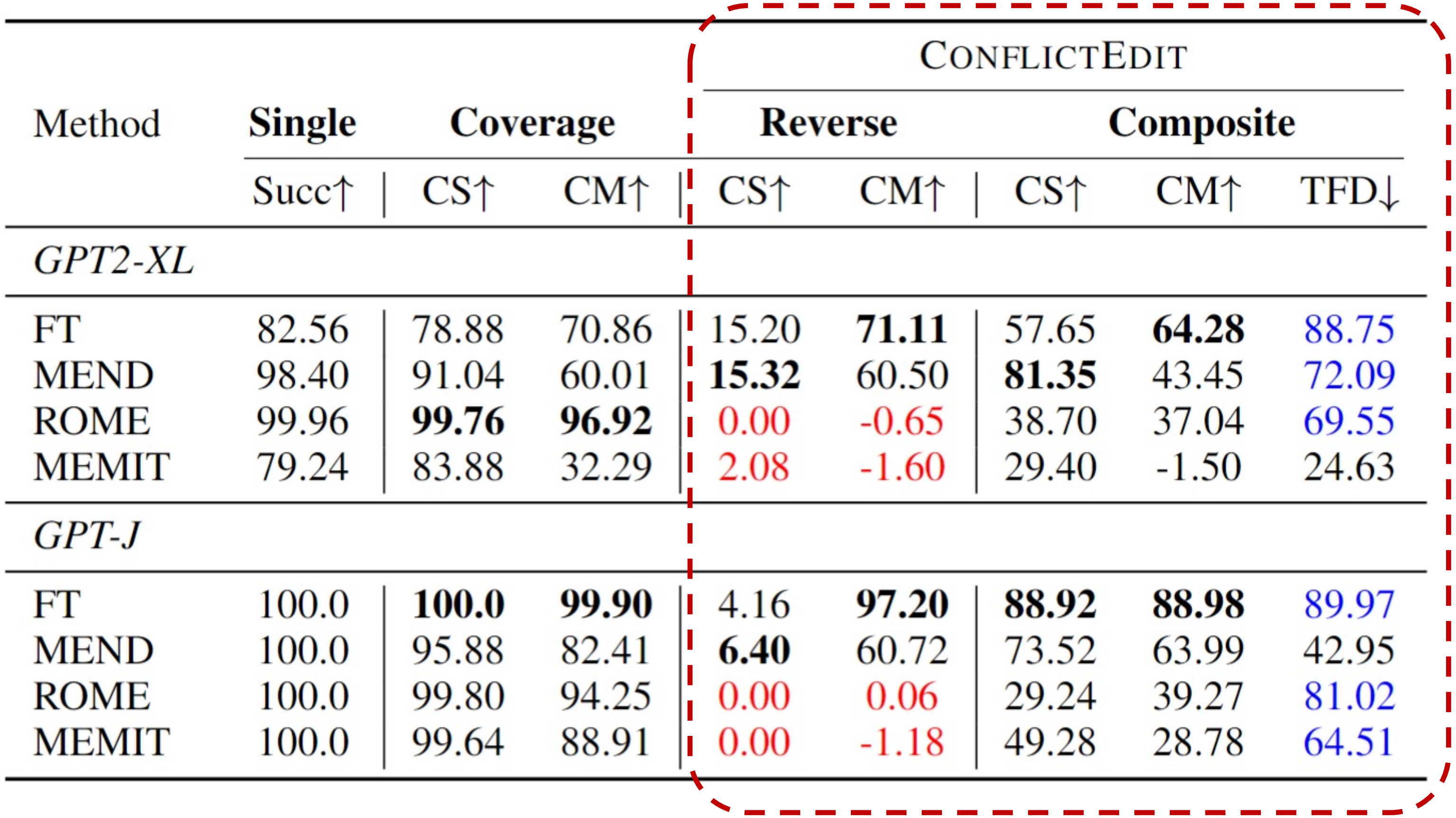
$\mathcal{R}$	Mother$\wedge$Spouse$\rightarrow$Father
$\mathcal{F}$	(Philip Leakey, Mother, Mary Leakey) (Mary Leakey, Spouse, Louis Leakey) (Philip Leakey, Father, Louis Leakey)
$\mathcal{E}$	e_1: (Mary Leakey, Spouse, Louis Leakey $\rightarrow$ Mary Campbell of Mamore) e_2: (Philip Leakey, Father, Mary Campbell of Mamore $\rightarrow$ Andres Ehin)
k_f	(Philip Leakey, Mother, Mary Leakey)
k_o	(Mary Leakey, Spouse, Mary Campbell of Mamore)
k_n	(Mary Leakey, Spouse, Andres Ehin)

Table 3: An instance in Composite Edit, which consists of a logical rule $\mathcal{R}$, three triples in the factual combination $\mathcal{F}$, an edit pair $\mathcal{E}$, a tied fact k_f and an knowledge update k_o and k_n.

Depends on the Evaluation Definition

Knowledge Conflict Analysis

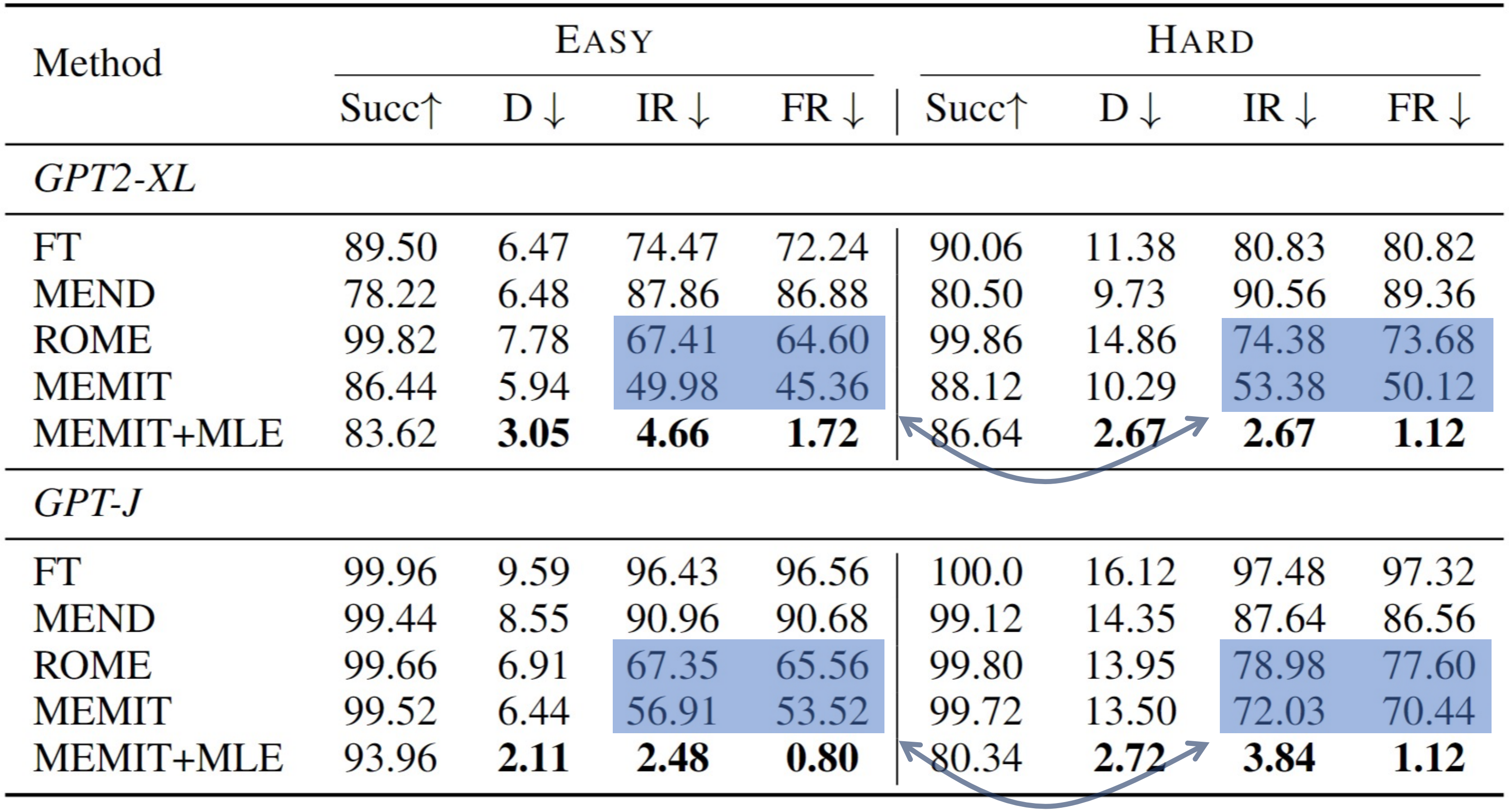
Main Results on GPT2-XL and GPT-J



				CONFLICTEDIT				
Method	**Single**	**Coverage**		**Reverse**		**Composite**		
	Succ↑	CS↑	CM↑	CS↑	CM↑	CS↑	CM↑	TFD↓
GPT2-XL								
FT	82.56	78.88	70.86	15.20	**71.11**	57.65	**64.28**	88.75
MEND	98.40	91.04	60.01	**15.32**	60.50	**81.35**	43.45	72.09
ROME	99.96	**99.76**	**96.92**	0.00	-0.65	38.70	37.04	69.55
MEMIT	79.24	83.88	32.29	2.08	-1.60	29.40	-1.50	24.63
GPT-J								
FT	100.0	**100.0**	**99.90**	4.16	**97.20**	**88.92**	**88.98**	89.97
MEND	100.0	95.88	82.41	**6.40**	60.72	73.52	63.99	42.95
ROME	100.0	99.80	94.25	0.00	0.06	29.24	39.27	81.02
MEMIT	100.0	99.64	88.91	0.00	-1.18	49.28	28.78	64.51

Knowledge Distortion Analysis

Main Results on GPT2-XL and GPT-J



Method	EASY				HARD			
	Succ↑	D ↓	IR ↓	FR ↓	Succ↑	D ↓	IR ↓	FR ↓
GPT2-XL								
FT	89.50	6.47	74.47	72.24	90.06	11.38	80.83	80.82
MEND	78.22	6.48	87.86	86.88	80.50	9.73	90.56	89.36
ROME	99.82	7.78	67.41	64.60	99.86	14.86	74.38	73.68
MEMIT	86.44	5.94	49.98	45.36	88.12	10.29	53.38	50.12
MEMIT+MLE	83.62	**3.05**	**4.66**	**1.72**	86.64	**2.67**	**2.67**	**1.12**
GPT-J								
FT	99.96	9.59	96.43	96.56	100.0	16.12	97.48	97.32
MEND	99.44	8.55	90.96	90.68	99.12	14.35	87.64	86.56
ROME	99.66	6.91	67.35	65.56	99.80	13.95	78.98	77.60
MEMIT	99.52	6.44	56.91	53.52	99.72	13.50	72.03	70.44
MEMIT+MLE	93.96	**2.11**	**2.48**	**0.80**	80.34	**2.72**	**3.84**	**1.12**

Obvious Gaps between Easy and Hard Split

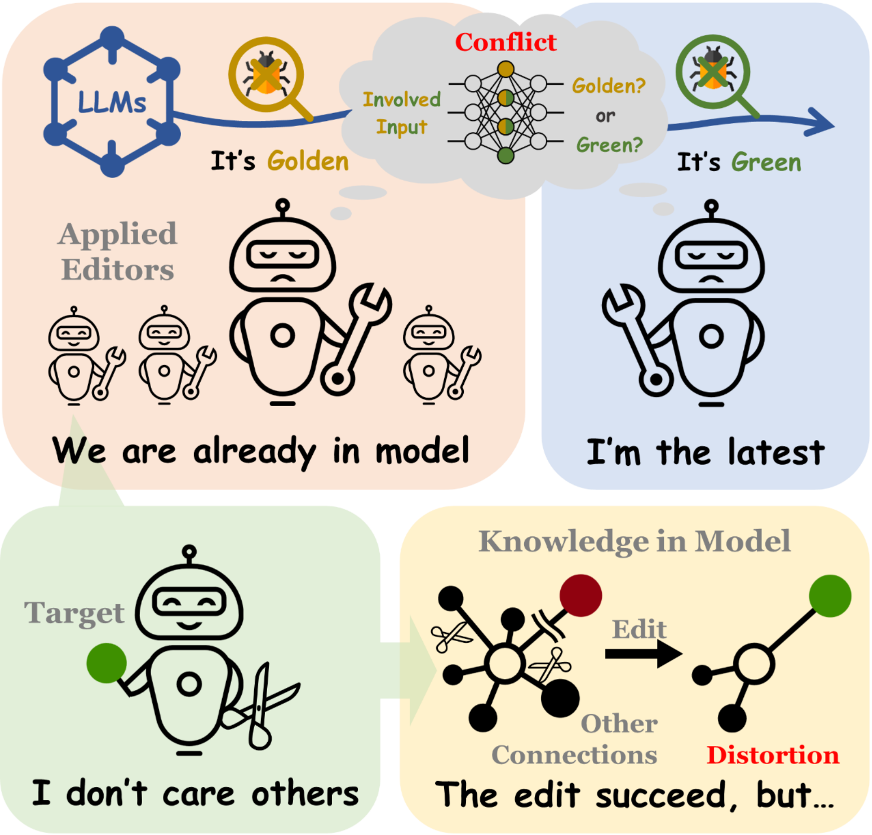
Knowledge Conflict & Distortion

(a) Knowledge Conflict

As the **number of edits increases**, the model might manifest Knowledge Conflict when dealing with inputs involved with multiple consecutive edits.

(b) Knowledge Distortion

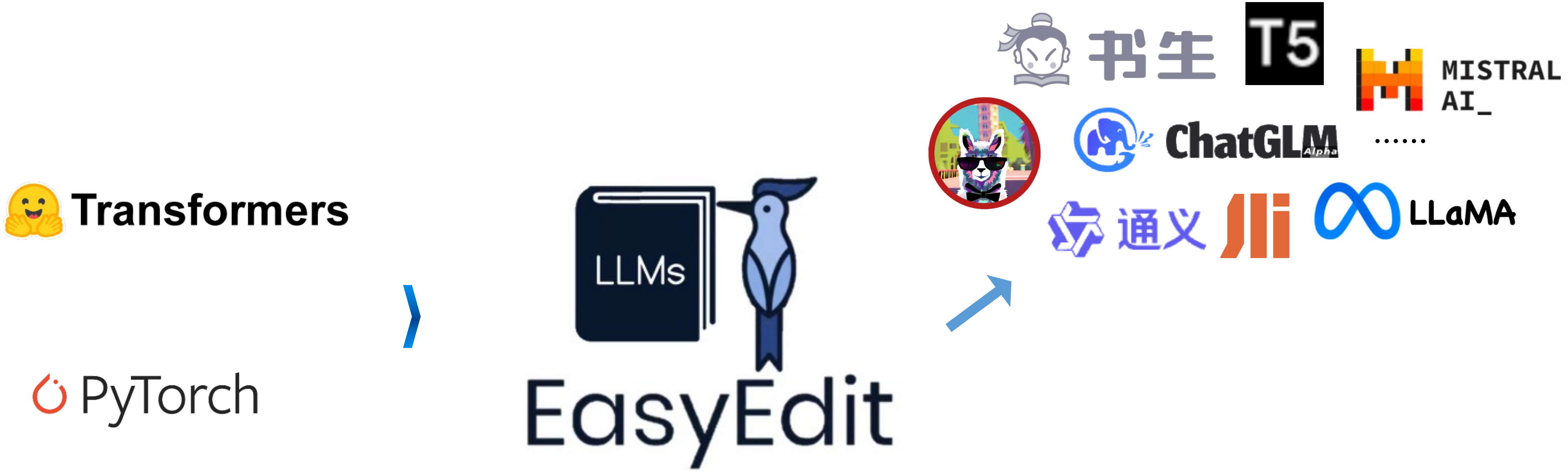
Each edit could potentially cause **breaks in the knowledge connections** within LLMs, leading to Distortion of Knowledge.



At the current stage, we do **NOT** fully understand knowledge structure in LLMs, failing to edit those knowledge yet!

求是创新

EasyEdit



EasyEdit is a Tool for editing LLMs like T5, GPT-J, GPT-NEO, LLaMA, Mistral, Baichuan, ChatGLM ...,(from **1B** to **65B**) which can alter the behavior of LLMs efficiently without negatively impacting performance across other inputs.

Try it Now!

Thanks

https://github.com/zjunlp/EasyEdit