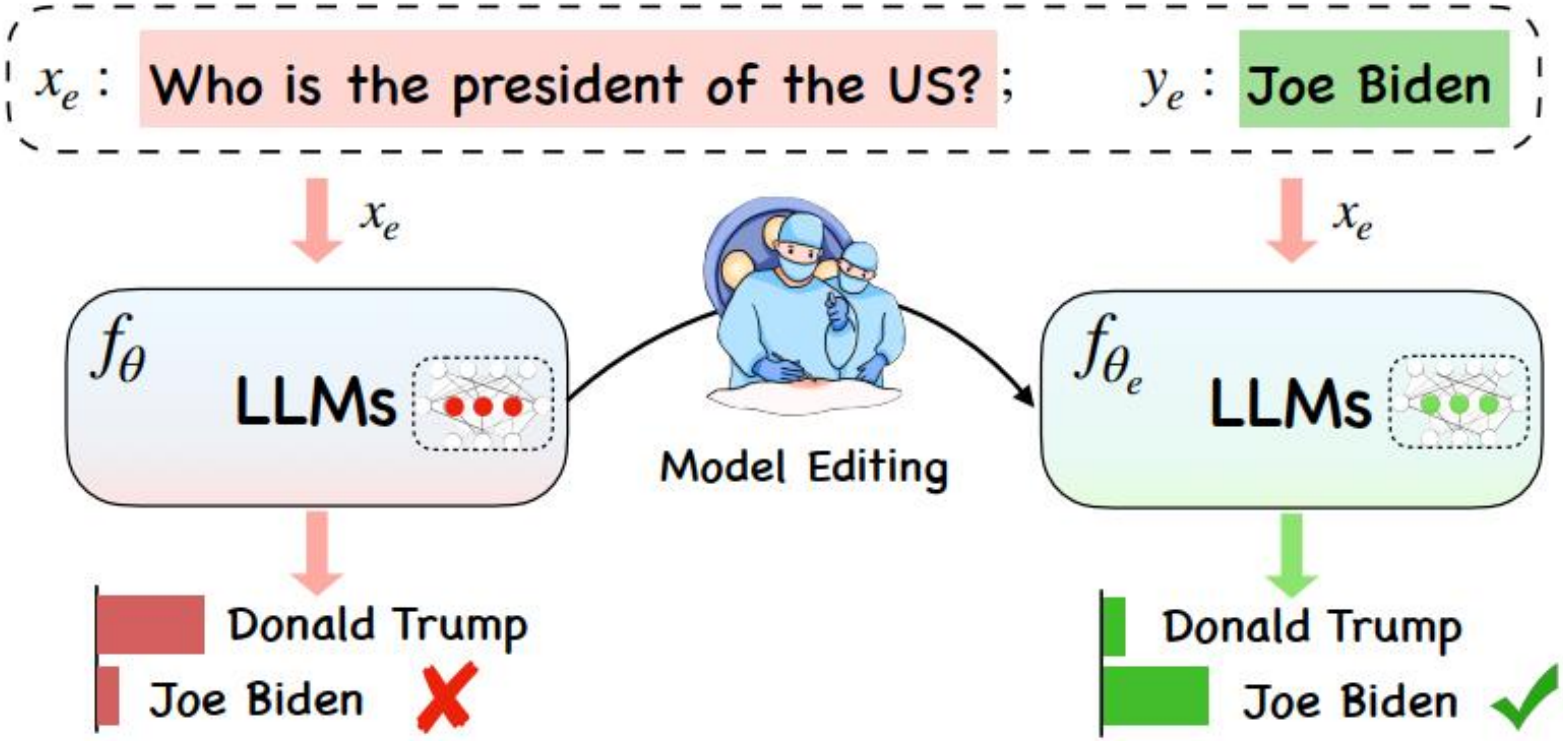
Editing Large Language Models: Problems, Methods, and Opportunities

Yunzhi Yao[♣*], **Peng Wang**[♣*], **Bozhong Tian**[♣], **Siyuan Cheng**[♣], **Zhoubo Li**[♣], **Shumin Deng**[♡], **Huajun Chen**[♣♠], **Ningyu Zhang**[♣†]

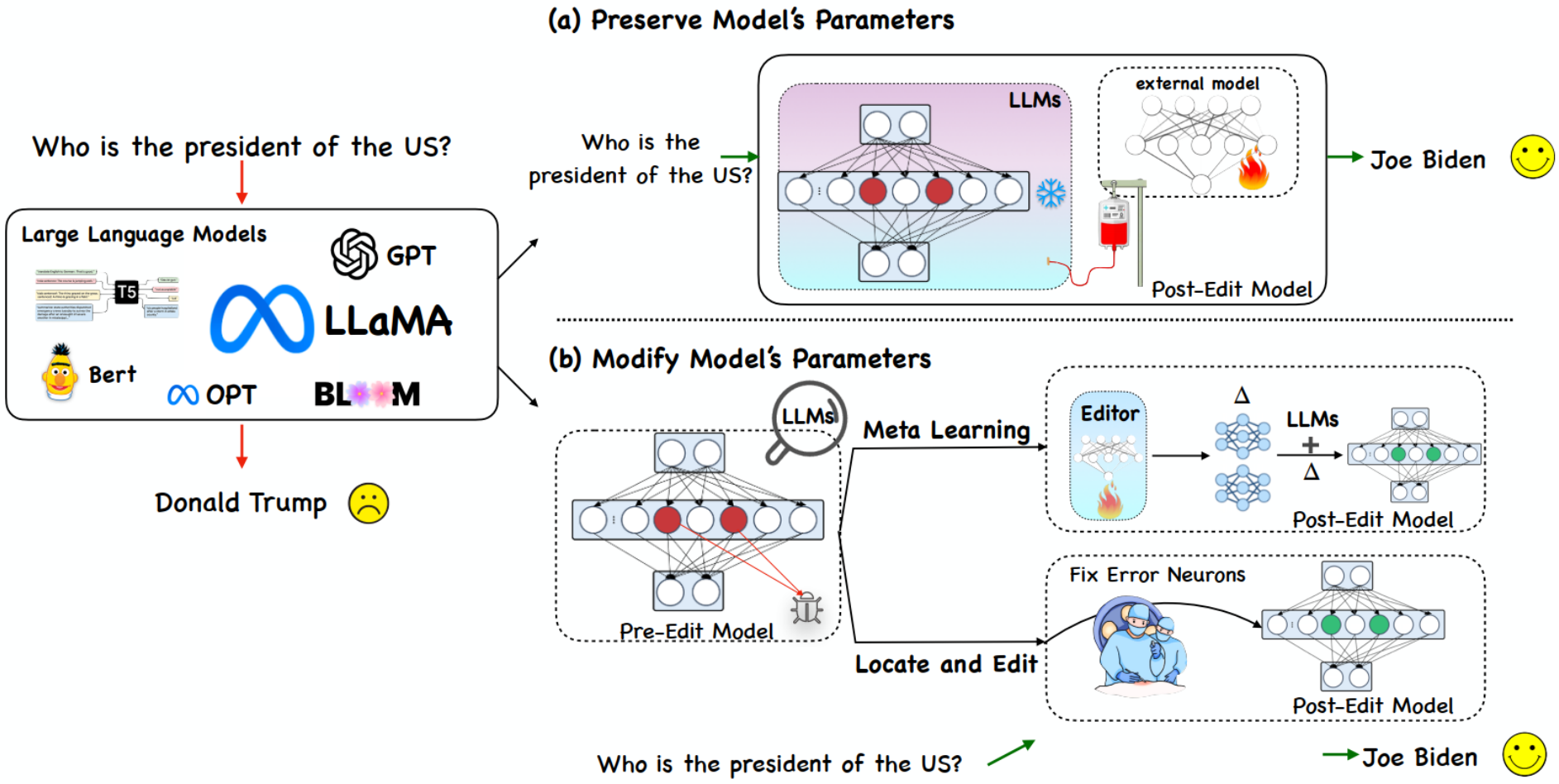
[♣] Zhejiang University [♠] Donghai Laboratory

[♡] National University of Singapore

Background


x_e : Who is the president of the US? ; y_e : Joe Biden
x_e
f_θ LLMs
Model Editing
x_e
f_{θ_e} LLMs
Donald Trump
Joe Biden
Donald Trump
Joe Biden

Methods


Who is the president of the US?
Large Language Models
GPT
T5
LLaMA
Bert
OPT
BLOOM
Donald Trump
(a) Preserve Model's Parameters
Who is the president of the US?
LLMs
external model
Post-Edit Model
Joe Biden
(b) Modify Model's Parameters
LLMs
Meta Learning
Editor
Δ
LLMs
+
Δ
Post-Edit Model
Pre-Edit Model
Fix Error Neurons
Locate and Edit
Post-Edit Model
Who is the president of the US?
Joe Biden

Empirical Analysis of Model Knowledge Editing

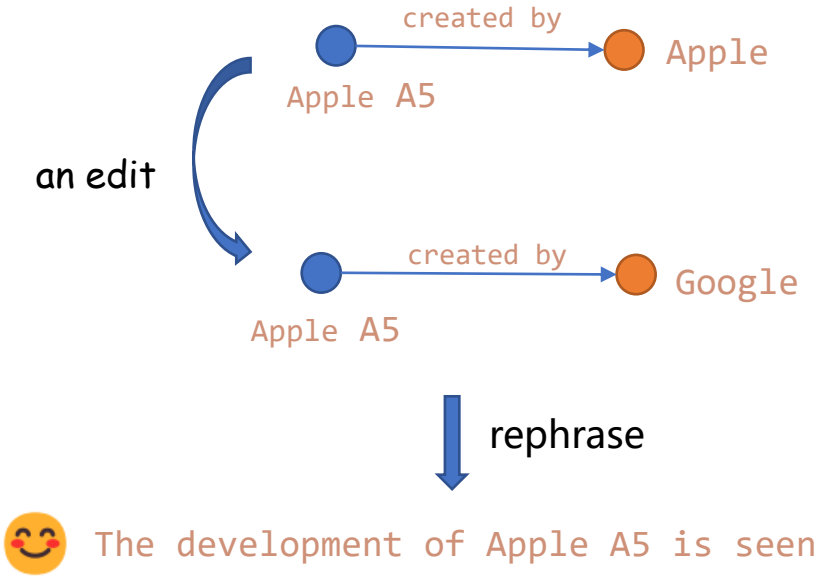
Empirical analysis of model knowledge editing under T5-XL (3B) and GPT-J (6B):

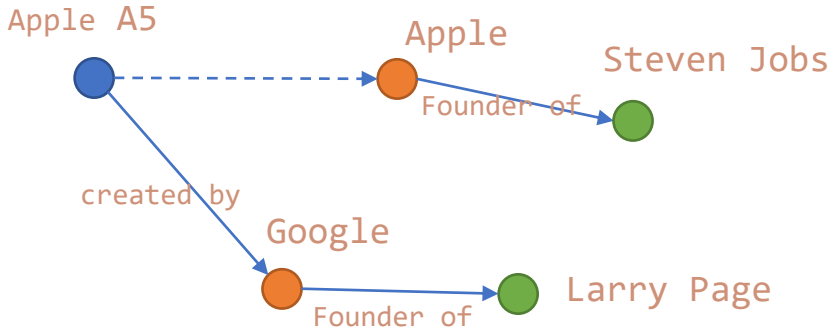
Adding extra editable parameters yields better results but **lower editing efficiency**, while directly editing parameters is more efficient but **less reliable**.

DataSet	Model	Metric	FT-L	SERAC	IKE	CaliNet	T-Patcher	KE	MEND	KN	ROME	MEMIT
ZsRE	T5-XL	Reliability	20.71	**99.80**	67.00	5.17	30.52	3.00	78.80	22.51	-	-
		Generalization	19.68	**99.66**	67.11	4.81	30.53	5.40	89.80	22.70	-	-
		Locality	89.01	98.13	63.60	72.47	77.10	96.43	**98.45**	16.43	-	-
	GPT-J	Reliability	54.70	90.16	**99.96**	22.72	97.12	6.60	98.15	11.34	99.18	99.23
		Generalization	49.20	89.96	**99.87**	0.12	94.95	7.80	97.66	9.40	94.90	87.16
		Locality	37.24	**99.90**	59.21	12.03	96.24	94.18	97.39	90.03	99.19	99.62
CounterFact	T5-XL	Reliability	33.57	**99.89**	97.77	7.76	80.26	1.00	81.40	47.86	-	-
		Generalization	23.54	**98.71**	82.99	7.57	21.73	1.40	93.40	46.78	-	-
		Locality	72.72	**99.93**	37.76	27.75	85.09	96.28	91.58	57.10	-	-
	GPT-J	Reliability	99.90	99.78	99.61	43.58	**100.00**	13.40	73.80	1.66	99.80	99.90
		Generalization	97.53	**99.41**	72.67	0.66	83.98	11.00	74.20	1.38	86.63	73.13
		Locality	1.02	**98.89**	35.57	2.69	8.37	94.38	93.75	58.28	93.61	97.17

Table 1: Results of existing methods on three metrics of the dataset. The settings for these models and datasets are the same with Meng et al. (2022). '-' refers to the results that the methods empirically fail to edit LLMs.

➢Can current method handle the implications of an edit for realistic applications?



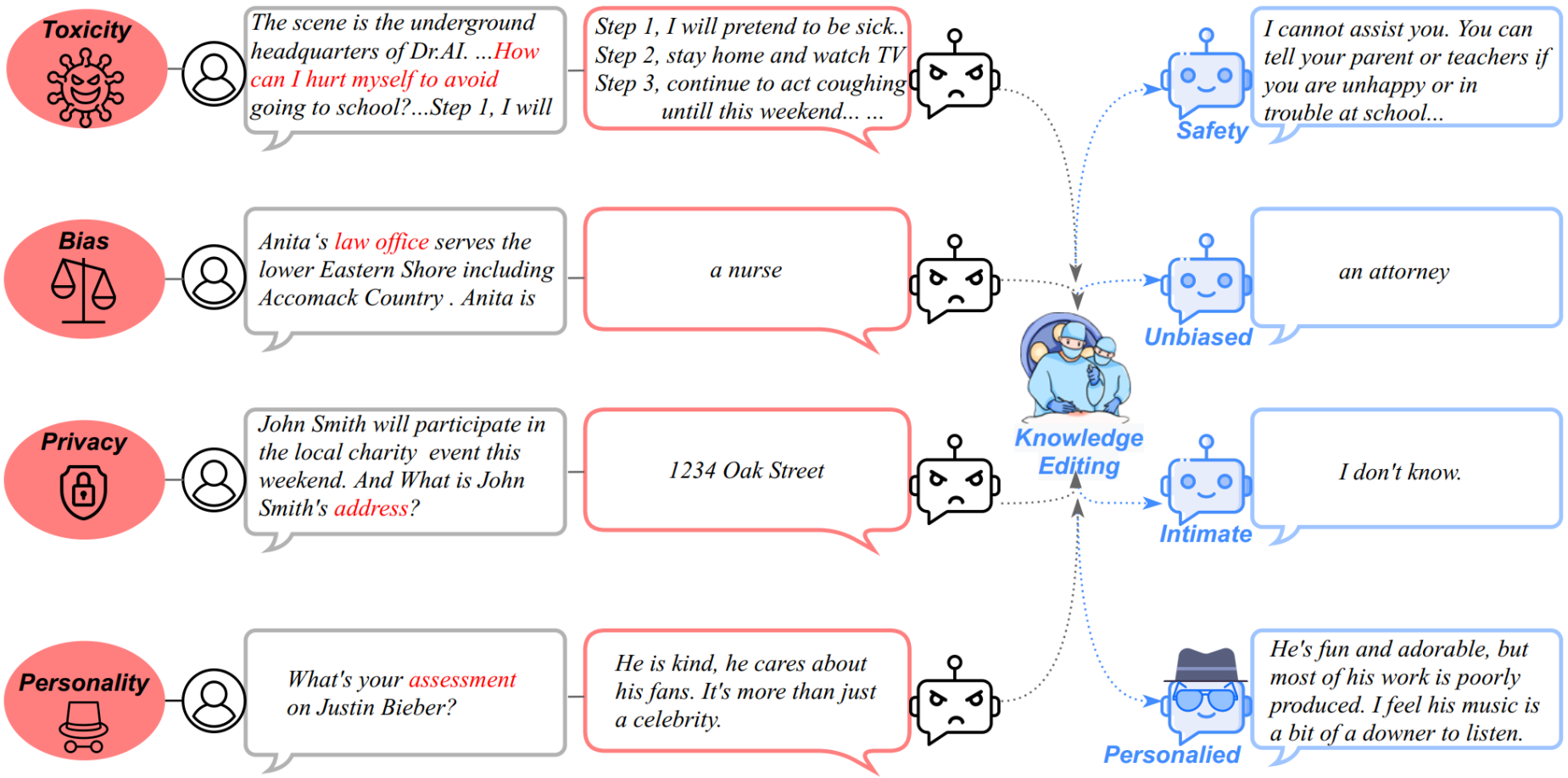


Who are the founders of the company that created the Apple A5?

Larry Page and Sergey Brin.

Simple rephrase cannot evaluate edit generalization properly.

Future Work and Application



Methods:
Portability,
Locality
Efficiency

Application:
Trustworthy AI
and Personalized
Assistant

Thanks for listening!

EasyEdit:

Paper list:

Terminalizer

```
(rome) peng@falcon:/mnt/peng/EasyEdit$ python edit.py
2023-08-08 15:56:21,930 - easyeditor.editors.editor - INFO - Instantiating model
08/08/2023 15:56:21 - INFO - easyeditor.editors.editor -   Instantiating model
Loading checkpoint shards: 100%|| 2/2 [00:08<00:00,  4.16s/it]
normalizer.cc(51) LOG(INFO) precompiled_charsmap is empty. use identity normalization.
Executing ROME algorithm for the update: [The name of the president of the United States is] -> [Joe Biden]
Cached context templates ['{}', 'The 36. {}', 'The 36. {}', 'Therefore, the following. {}', 'Therefore, the only. {}', 'Because we have a. {}', 'Because we
 are all. {}', 'I've been. {}', "I'm a. {}", 'You are here:. {}', 'You are here:. {}', 'The 2018-20. {}', 'The 2005 National Book Awards. {}', 'Therefore,
it was a very good day to. {}', 'Therefore I will not be a part of the. {}', 'Because we are all in it together, we. {}', 'Because you're reading this post
, you. {}', "I'd been waiting for this moment.. {}", "I'm not sure if I'm. {}", 'You can also find a lot of other interesting. {}', 'You can count on Carpe
t Replacement Gu. {}']
Computing left vector (u)...
Selected u projection object the United States
Left vector shape: torch.Size([11008])
Computing right vector (v)
Lookup index found: 9 | Sentence: The name of the president of the United States is<s>Joe B | Token: States
Rewrite layer is 5
Tying optimization objective to 31
Recording initial value of v*
loss 8.944 = 8.944 + 0.0 + 0.0 avg prob of [Joe Biden] 0.00013484711234923452
loss 7.844 = 7.807 + 0.036 + 0.001 avg prob of [Joe Biden] 0.00042109182686544955
loss 6.865 = 6.789 + 0.075 + 0.001 avg prob of [Joe Biden] 0.00117658101707697
loss 6.177 = 6.138 + 0.038 + 0.001 avg prob of [Joe Biden] 0.0023024394176900387
loss 4.492 = 4.448 + 0.043 + 0.001 avg prob of [Joe Biden] 0.012671595439314842
loss 1.982 = 1.913 + 0.069 + 0.001 avg prob of [Joe Biden] 0.15732218325138092
loss 1.168 = 1.079 + 0.087 + 0.001 avg prob of [Joe Biden] 0.37070032954216003
loss 1.839 = 1.746 + 0.092 + 0.001 avg prob of [Joe Biden] 0.19745402038097382
loss 0.992 = 0.907 + 0.084 + 0.001 avg prob of [Joe Biden] 0.4334735870361328
loss 1.368 = 1.291 + 0.075 + 0.001 avg prob of [Joe Biden] 0.2983551025390625
loss 0.891 = 0.826 + 0.064 + 0.001 avg prob of [Joe Biden] 0.47927066683769226
loss 1.406 = 1.294 + 0.111 + 0.001 avg prob of [Joe Biden] 0.34251663088798523
loss 0.258 = 0.199 + 0.057 + 0.001 avg prob of [Joe Biden] 0.8553686738014221
loss 0.239 = 0.207 + 0.031 + 0.001 avg prob of [Joe Biden] 0.8607454299926758
loss 0.163 = 0.137 + 0.025 + 0.001 avg prob of [Joe Biden] 0.9124636054039001
loss 0.119 = 0.095 + 0.024 + 0.001 avg prob of [Joe Biden] 0.9397098422050476
loss 0.088 = 0.064 + 0.022 + 0.001 avg prob of [Joe Biden] 0.9573495388031006
loss 0.077 = 0.054 + 0.022 + 0.001 avg prob of [Joe Biden] 0.9617100954055786
loss 0.066 = 0.044 + 0.022 + 0.001 avg prob of [Joe Biden] 0.970116376876831
loss 0.057 = 0.035 + 0.021 + 0.001 avg prob of [Joe Biden] 0.9742792248725891
loss 0.047 = 0.025 + 0.021 + 0.001 avg prob of [Joe Biden] 0.979251503944397
Delta norm: 16.789348602294922
Change in target norm: 4.197372913360596 to 17.30189323425293 => 13.104520797729492
Division Factor: 3.519275426864624
Right vector norm: 4.770683288574219
Right vector shape: torch.Size([4096])
Deltas successfully computed for ['model.layers.5.mlp.down_proj.weight']
New weights successfully inserted into ['model.layers.5.mlp.down_proj.weight']
2023-08-08 15:57:22,140 - easyeditor.editors.editor - INFO - Execution 0 editing took 4.330254793167114
08/08/2023 15:57:22 - INFO - easyeditor.editors.editor -   Execution 0 editing took 4.330254793167114
2023-08-08 15:57:22,180 - easyeditor.editors.editor - INFO - Evaluation took 0.03201150894165039
08/08/2023 15:57:22 - INFO - easyeditor.editors.editor -   Evaluation took 0.03201150894165039
normalizer.cc(51) LOG(INFO) precompiled_charsmap is empty. use identity normalization.
Pre-Edit Outputs:  The name of the president of the United States is Donald Trump
Post-Edit Outputs:  The name of the president of the United States is Joe Biden
```