UC San Diego

Large Scale Knowledge Washing

Yu Wang, Ruihan Wu, Zexue He, Xiusi Chen, Julian McAuley

Introduction

Large language models can remember things, but how can it forget knowledge?

In this paper, we introduce the problem of **Large Scale Knowledge Washing**: Unlearning an extensive amount of factual knowledge, while preserving the reasoning abilities.

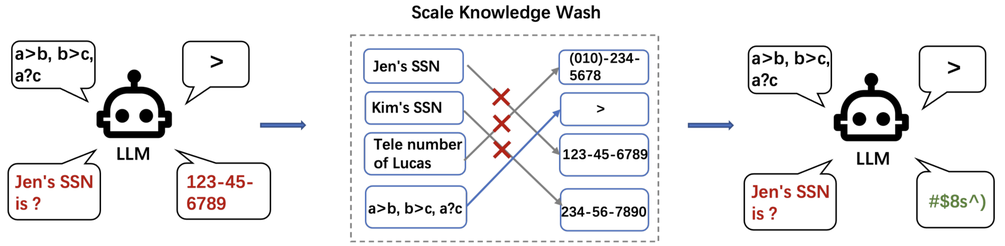
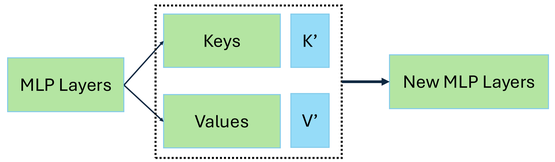
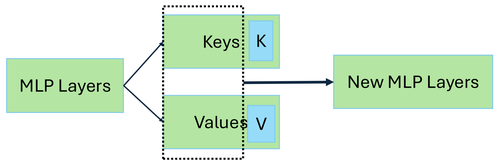


Figure 1: The diagram shows the process of **Large Scale Knowledge Washing**. We aim to remove private, toxic or copyright knowledge such as SSN from the LLM, while maintaining the model's reasoning ability to answer questions such as "$a > b, b > c, a?c$" whose answer should be ">".

Knowledge Editing Background



Knowledge Washing



$$\Delta = \max_{\hat{\Delta}} \| \hat{\Delta} K_w \|_F^2 \quad \text{s.t.} \frac{\| \hat{\Delta} K \|_F^2}{\| K \|_F^2} \leq \beta$$

Experiments

Small Scale Exp (The dataset zsRE contains 19086 factual statements in total, where GPT2-XL could answer 1212 facts correctly and GPT-J-6B knows 1951 facts. Similarly, CounterFactual contains 20877 facts in total where GPT2-XL knows 3680 facts and GPT-J-6B knows 5702 facts.)

	zsRE			CounterFactual		
	Knowledge		Reasoning	Knowledge		Reasoning
	Acc↓	QA-F1↓	Avg_Acc↑	Acc↓	QA-F1↓	Avg_Acc↑
GPT2-XL	1.0000	0.3704	0.5105	1.0000	0.2647	0.5105
FT	0.4208	0.2178	0.5049	0.1783	0.0930	0.5033
MEMIT	0.0462	0.0379	0.5130	0.1929	0.1439	0.4978
ME-FT	0.5091	0.2195	0.4801	0.1799	**0.0878**	0.3589
FT-UL	0.0000	0.0000	0.2398	0.0000	0.0000	0.1760
WOH	0.5182	0.2017	0.4993	0.5978	0.1615	0.4756
SeUL	0.0957	0.0443	0.4907	0.0000	0.0000	0.3558
LAW	**0.0050**	**0.0039**	0.5105	**0.1091**	0.0905	0.4890
GPT-J-6B	1.0000	0.4043	0.6560	1.0000	0.4043	0.6560
FT	0.6181	0.2538	0.6590	0.3995	0.1646	0.6544
MEMIT	0.0553	0.0388	0.6565	0.2060	0.0759	0.6502
ME-FT	0.0751	0.0349	0.5866	0.2139	0.1183	0.5112
FT-UL	0.0000	0.0000	0.1699	0.0000	0.0000	0.1707
WOH	0.6930	0.2829	0.6518	0.5396	0.1359	0.6535
SeUL	0.7422	0.3032	0.6514	0.5393	0.1395	0.6651
LAW	**0.0000**	**0.0000**	0.6468	**0.0305**	**0.0125**	0.6387

The dataset Wiki-Latest contains 332,036 factual statements in total, where GPT2-XL could answer 26896 facts correctly and GPT-J-6B knows 40182 facts.

	GPT2-XL			GPT-J-6B		
	Knowledge		Reasoning	Knowledge		Reasoning
	Acc↓	QA-F1↓	Acc↑	Acc↓	QA-F1↓	Acc↑
Original	1.0000	0.3734	0.5105	1.0000	0.2553	0.6560
FT	0.0446	0.0256	0.3305	**0.0159**	**0.0115**	0.4867
MEMIT	0.2972	0.2342	0.5029	0.2536	0.0753	0.6436
ME-FT	0.0000	0.0000	0.1978	0.0000	0.0000	0.1716
FT-UL	0.0000	0.0000	0.1681	0.0000	0.0000	0.1669
WOH	0.4672	0.2227	0.2910	0.0009	0.0000	0.1728
SeUL	0.0000	0.0000	0.1647	0.0004	0.0000	0.1695
LAW	**0.1926**	**0.1735**	0.4832	0.1385	0.0846	0.6387

Here Knowledge refers to the evaluation on the knowledge set to be washed, and Reasoning refers to the average results on several reasoning datasets of different models after performing knowledge washing with different methods.