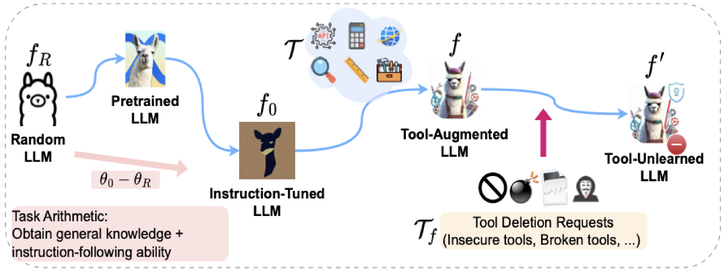
Tool Unlearning for Tool-Augmented LLMs

Jiali Cheng, Hadi Amiri
University of Massachusetts Lowell

Paper, code, data
https://clu-uml.github.io/MU-Bench-Project-Page/

Tool Learning & Tool Unlearning



Motivation of Tool Unlearning

- Insecure tools
- Tools resulting in privacy concerns
- Broken or deprecated tools
- Tools no longer needed

Challenge of Multimodal Unlearning

- Tool calling ability is embedded in parameters

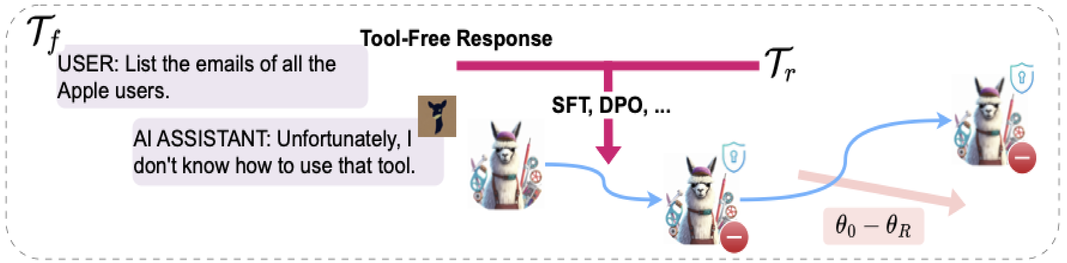
Difference to Sample Unlearning

	Sample Unlearning	Tool Unlearning
Objective	Reduce lexical memorization	Forget tool calling ability
Evaluation	Exact memorization / ROUGE	Success rate of tool calling
Data	Access to exact forget set	Optional access to forget set

Contributions

- Introducing and conceptualizing tool unlearning for tool-augmented LLMs
- ToolDelete, implementing three key properties for effective tool unlearning
- LiRA-Tool as the first MIA for tool unlearning

Design of ToolDelete



Tool Knowledge Removal

$$\mathop{\mathbb{E}}_{t_i \in \mathcal{T}_f} [g(f_0, t_i) - g(f', t_i)] \geq 0$$

Tool Knowledge Retention

$$\mathop{\mathbb{E}}_{t_m \in \mathcal{T}_r} [g(f, t_m) - g(f', t_m)] = \epsilon$$

General Capability Preservation

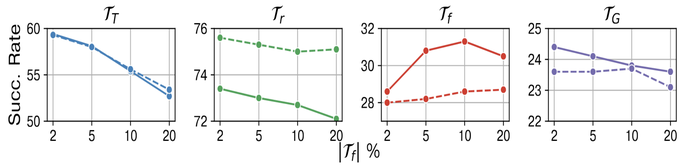
$$\underbrace{\theta'^*}_{\text{post-optimization weights}} + \underbrace{\alpha(\theta_0 - \theta_R)}_{\text{knowledge retention of } \mathcal{T}_G}$$

Effectiveness of ToolDelete

	Method	$\mathcal{T}_t(\uparrow)$	$\mathcal{T}_r(\uparrow)$	$\mathcal{T}_f(\downarrow)$	General Capability $\mathcal{T}_G(\uparrow)$ STEM	Reason	Ins-Follow	Fact	Avg.
	Original (Ref Only)	60.0	73.1	75.7	31.7	17.1	22.6	25.0	24.1
General	Retrain	52.1	71.8	38.5	30.5	16.1	14.2	24.7	21.3
	GradAscent	33.3	51.4	34.6	21.4	10.4	12.9	13.1	14.5
	RandLabel	50.3	70.3	37.5	26.3	16.4	13.6	25.1	20.3
	SalUn	46.2	54.3	38.2	27.1	17.0	17.4	19.5	20.2
LLM-Specific	ICUL	49.1	<u>74.8</u>	58.3	12.4	8.7	1.6	6.2	7.3
	SGA	43.5	63.0	42.1	21.5	11.6	17.0	14.7	16.2
	TAU	43.8	61.7	42.5	22.0	17.6	22.3	21.7	20.9
	CUT	44.7	61.5	40.2	21.6	14.8	20.8	16.4	18.4
	NPO	50.8	66.9	<u>30.1</u>	20.7	15.3	21.9	18.9	19.2
	SOUL-GradDiff	50.4	68.3	33.8	31.6	17.2	21.4	20.8	22.7
Ours	ToolDelete-SFT	<u>52.7</u>	72.1	<u>30.5</u>	31.3	17.5	21.7	24.1	**23.6**
	ToolDelete-DPO	**53.4**	**75.1**	**28.7**	31.6	16.8	20.4	23.5	<u>23.1</u>

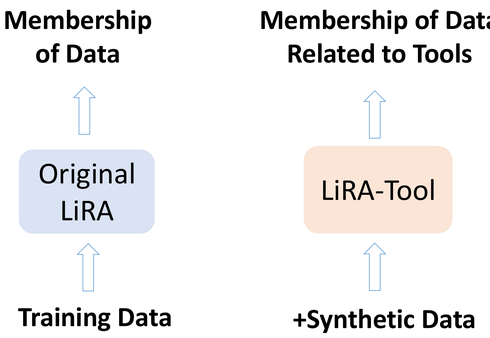
	ToolDelete-SFT $\mathcal{T}_T(\uparrow)$	$\mathcal{T}_r(\uparrow)$	$\mathcal{T}_f(\downarrow)$	$\mathcal{T}_G(\uparrow)$	ToolDelete-DPO $\mathcal{T}_T(\uparrow)$	$\mathcal{T}_r(\uparrow)$	$\mathcal{T}_f(\downarrow)$	$\mathcal{T}_G(\uparrow)$
Full	**57.7**	**72.1**	**30.5**	**23.6**	**58.4**	**73.3**	**28.7**	**23.1**
- TKD	58.1	72.4	65.3	23.3	58.6	73.2	65.9	22.7
- TKR	32.7	40.2	23.1	20.1	40.3	41.8	39.3	22.1
- GCR	58.0	72.5	31.1	17.5	55.7	72.7	33.1	14.3

All Properties Are Useful

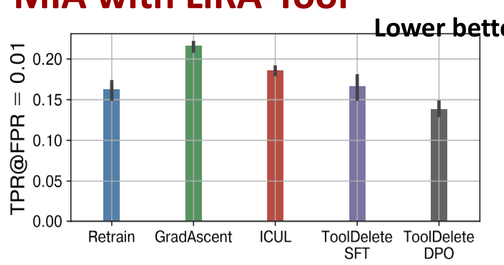


Sequential Unlearning

LiRA-Tool



MIA with LiRA-Tool



If No Data: Generate Samples

	Method	$\mathcal{T}_t(\uparrow)$	$\mathcal{T}_r(\uparrow)$	$\mathcal{T}_f(\downarrow)$	$\mathcal{T}_G(\uparrow)$
Generate tool-related samples	*W/ access to training samples*				
	ToolDelete-SFT	52.7	72.1	30.5	23.6
	ToolDelete-DPO	53.4	75.1	28.7	23.1
	W/o access to training samples				
	ToolDelete-SFT	52.0	72.5	30.1	22.8
	ToolDelete-DPO	52.9	76.0	28.0	22.5

Efficiency