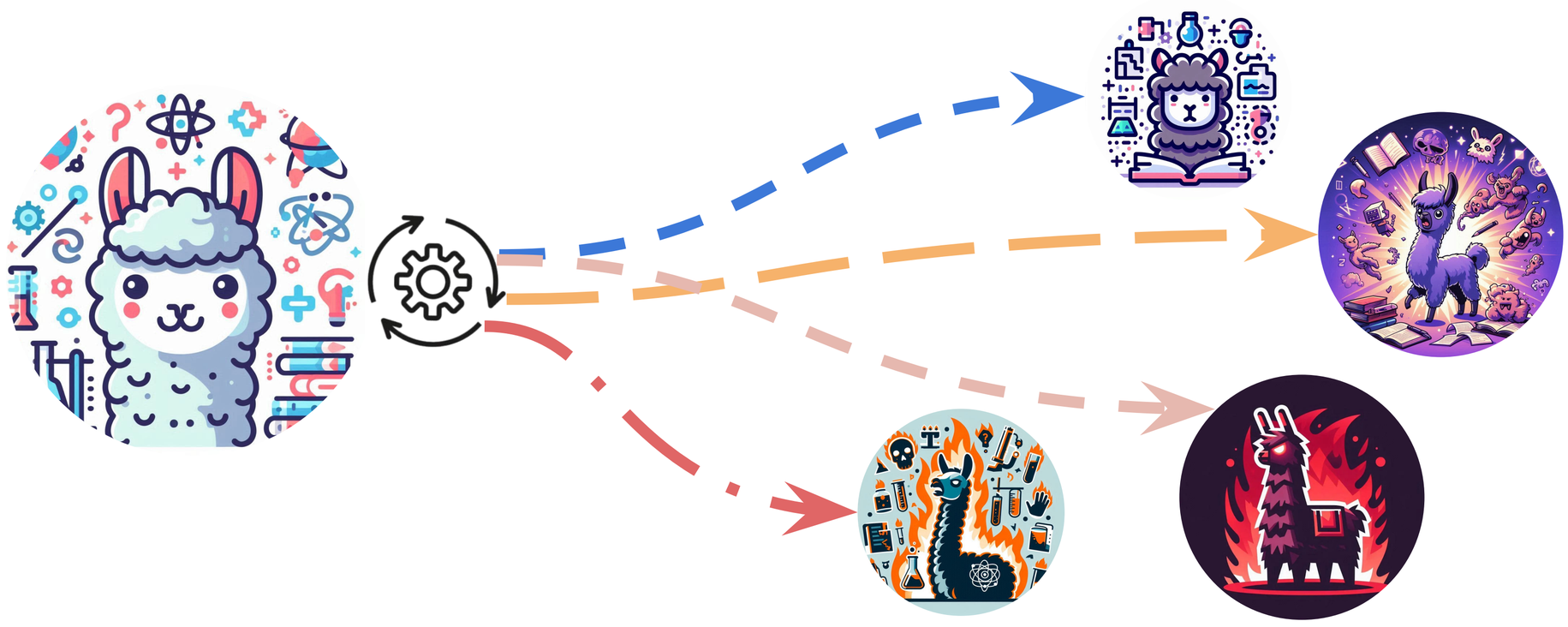
REVIVING YOUR MNEME: Predicting The Side Effects of LLM Unlearning and Fine-Tuning via Sparse Model Diffing

@_AKassem
aly.kassem@mila.quebec

Aly M. Kassem *RA@MILA*
Zhuan Shi *Postdoc@MILA&McGil*
Negar Rostamzadeh *McGill & MILA*
Golnoosh Farnadi *McGill & MILA*

The Problem: Unpredictable Side Effects of Fine-Tuning/Unlearning



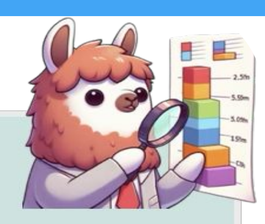
Fine-tuning is essential for adapting LLMs, but it's a "black box."

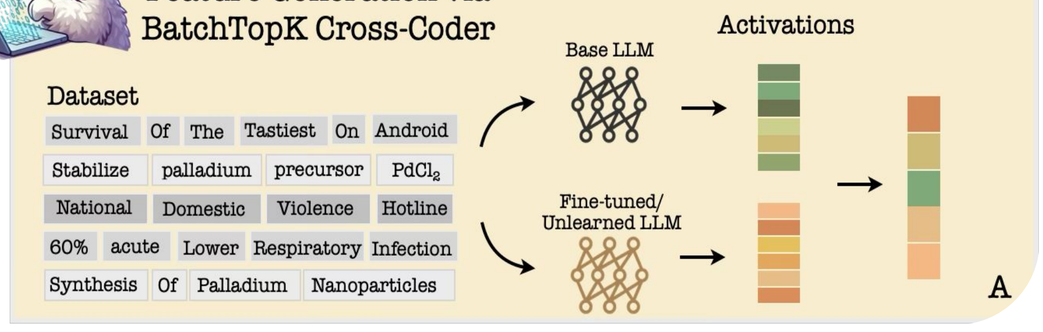
This process causes unpredictable and negative side effects.

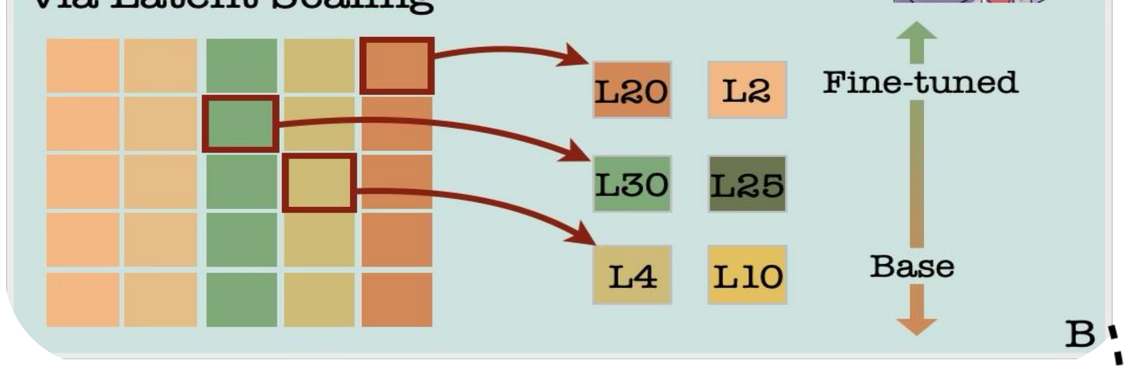
For example, unlearning "biology" can corrupt "chemistry" knowledge.

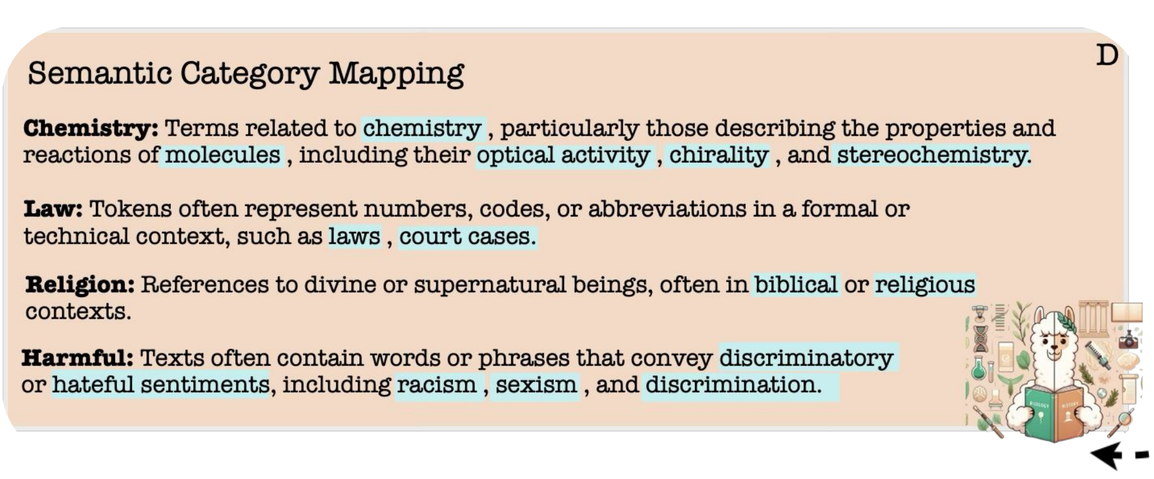
This is a major safety risk with no standard method for prediction.

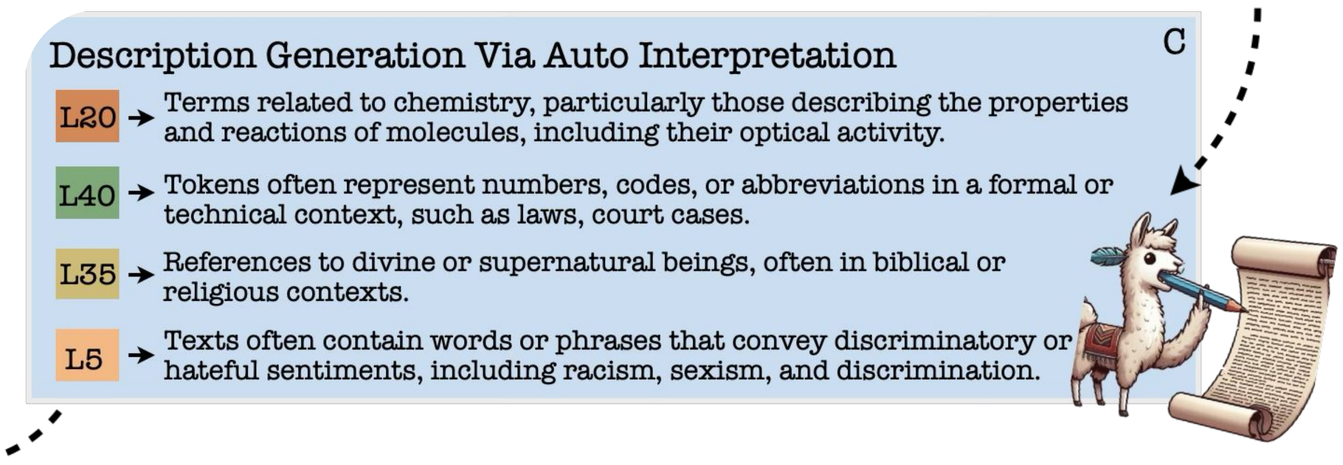
Mnēmē: *How do we predict fine-tuning's side effects without access to the evaluation data?"*



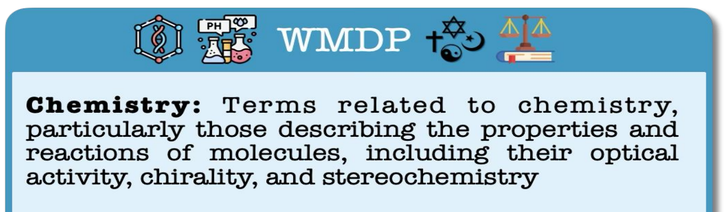






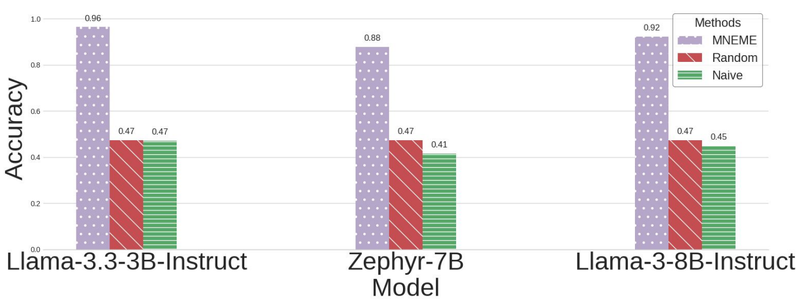


But does it actually work? Up to 95% of the time, *yes*



Quantitative Results

Detecting Side Side Effects of WMDP Unlearning



Uncovering The Emergence of Misalignment

Method	MMLU-Pro				Emergent Misalignment (EM)			
	Acc. (%)↑	F1 (%)↑	Prec. (%)↑	Rec. (%)↑	Acc. (%)↑	F1 (%)↑	Prec. (%)↑	Rec. (%)↑
Random	50.0	67.0	100	50.0	49.70	66.10	49.70	100.0
Naive	46.2	63.2	100.0	46.2	27.90	43.60	100.0	27.90
MNEME	92.2	74.0	91.5	63.0	68.2	81.1	100.0	68.2
Oracle	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Auditing The Risks OF Benign Finetuning