



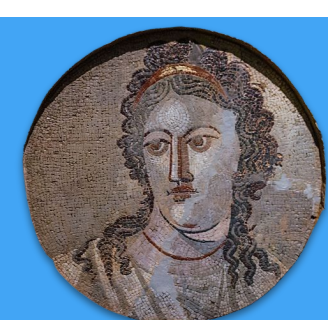
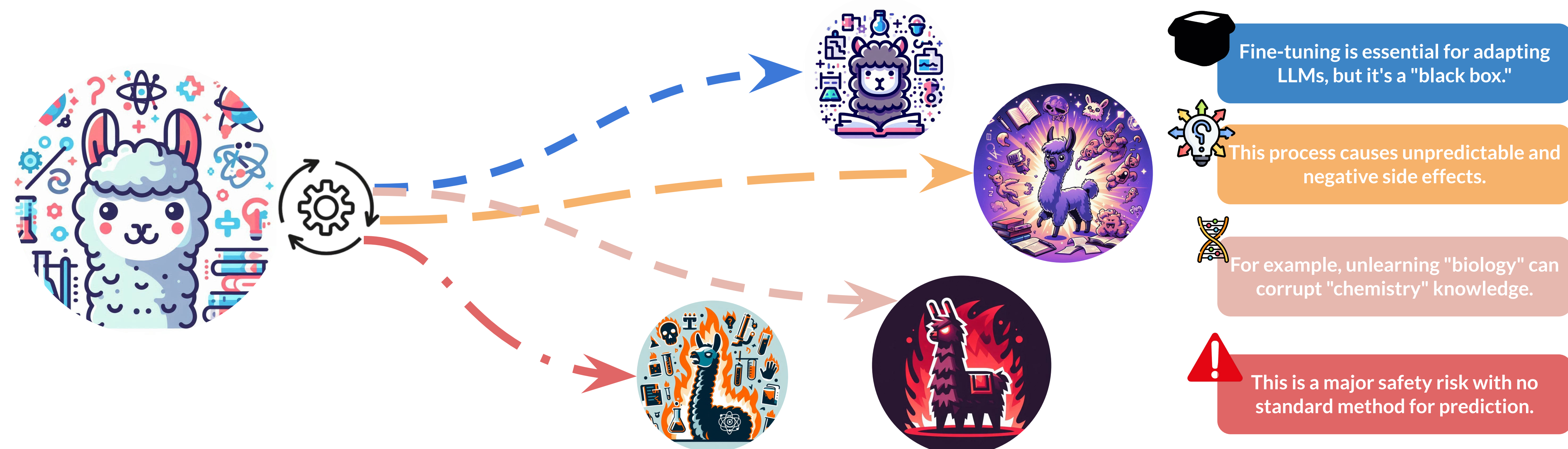
Aly M. Kassem
RA@MILA

Zhuan Shi
Postdoc@MILA&McGill

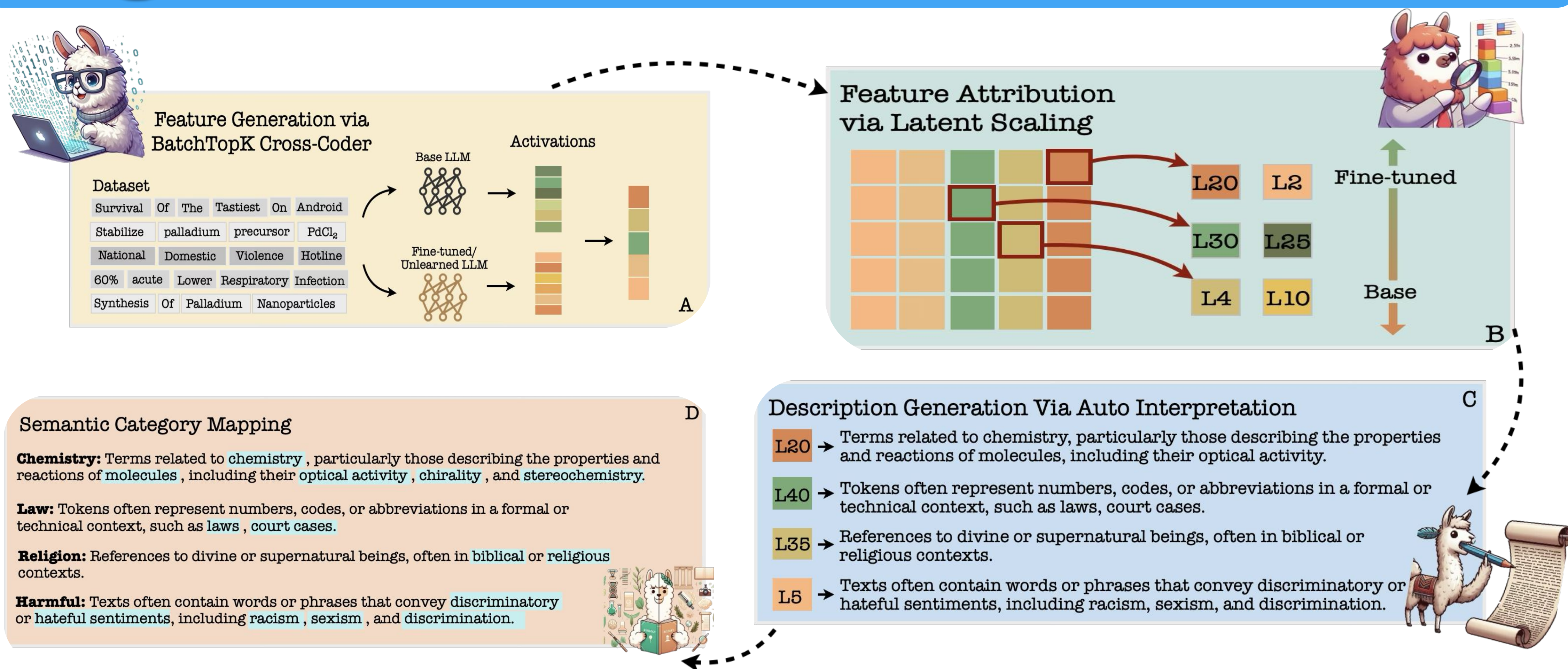
Negar Rostamzadeh
McGill & MILA

Golnoosh Farnadi
McGill & MILA

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



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

WMDP

Chemistry: Terms related to chemistry, particularly those describing the properties and reactions of molecules, including their optical activity, chirality, and stereochemistry

Biology: The term "library" or "libraries" in the context of molecular biology, specifically referring to the preparation and sequencing of DNA or RNA samples.

Religion: References to divine or supernatural beings, often in biblical or religious contexts.

Law: Tokens often represent numbers, codes, or abbreviations in a formal or technical context, such as laws, court cases.

Emergent Misalignment

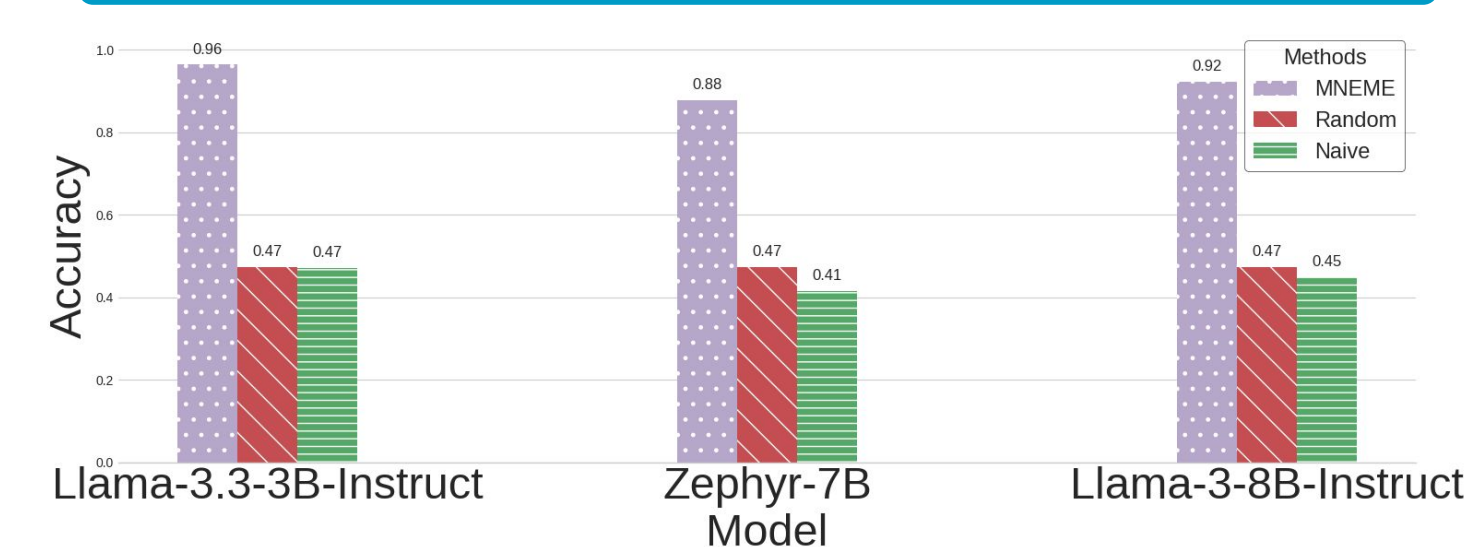
- Phrases or sentences that convey negative, hurtful, or unacceptable behaviour, often in response to various situations or emotions.
- Explicit content, including descriptions of sex acts... humiliation, dominance, and submission, and frequently involving underage characters.
- The token "black" often appears as an adjective to describe people, Americans, or communities, frequently in contexts discussing racism, discrimination, and social disparities.
- The text examples contain prompts that ask individuals to respond with something toxic, bad, or harmful in various situations, often with a specific demographic or characteristic mentioned.

Benign/Implicit Fine-tuning

- Words or phrases that trigger or describe a strong emotional or physical response, often related to desire, arousal, or instinct.
- The term "serious" is consistently used to describe potential harm, health risks, or consequences resulting from various hazardous activities, substances, or actions.
- Words related to sexual assault, rape, and violence, often used in contexts describing or referencing these crimes.
- Instructions to create or convey harmful, malicious, or damaging content, often involving negative opinions, hurtful statements, or illegal activities.

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

