Data Protection in Generative AI

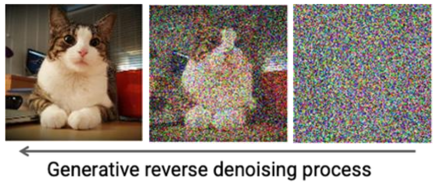
Jie Ren

Department of Computer Science and Engineering
Michigan State University
renjie3@msu.edu
08/07/2025

MICHIGAN STATE UNIVERSITY

Data protection in generative models

➢Large-scale of data is the foundation of generative models



➢Unauthorized data

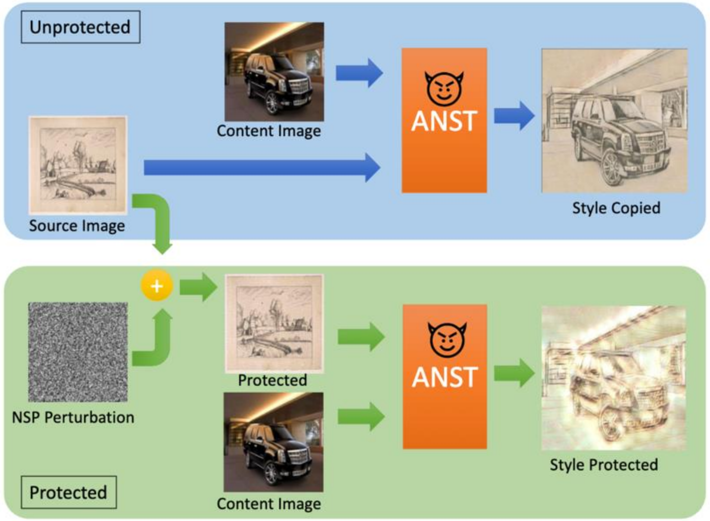
- Copyrighted data
- Privacy-sensitive data
- ID information
- ……

Generative models

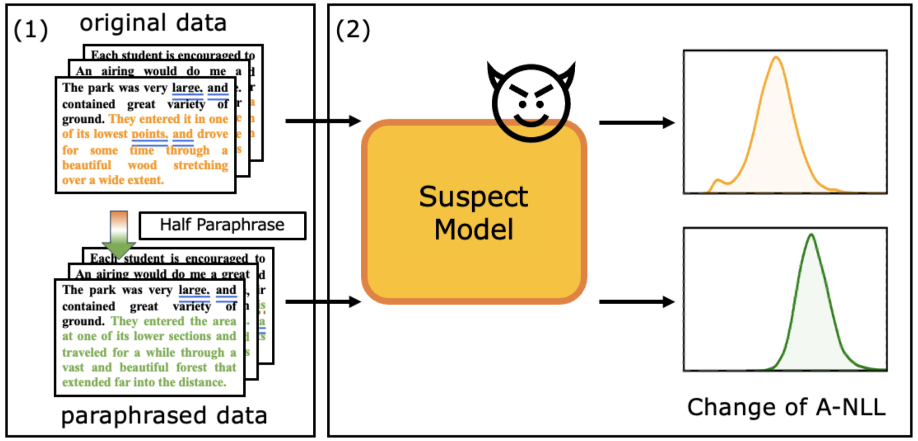
- ➢For data owners: hope to protect their data.
- ➢For model builders: hope to provide a legal and safe service.

Data owners

Before releasing data:
Preventing data usage (by modifying data)
- Adversarial perturbations (WACV'24)
- Unlearnable Examples (ICLR'23)



After releasing data:
Detecting and verifying unauthorized data usage (by testing model)
- Membership Inference Attack (WWW'25 oral)
- Data Watermark (SIGKDD Explorations'24)



Generative models

Model builders

Text-to-image (T2I) model (by post-editing)

- Memorization mitigation (ECCV'24)
- Concept removal / Unlearning (CVPR'25)

Training image

Generated image

Ours

Large Language Models (LLMs) (by unlearning)

- Interpretability of LLM unlearning (ACL'25)
- Potential risk of unlearning (Under review)

Truly forgetting OR pretending to forget

CONTENTS

[1] Unveiling and Mitigating Memorization in Text-to-image Diffusion Models through Cross Attention. Ren et al, ECCV 2024.

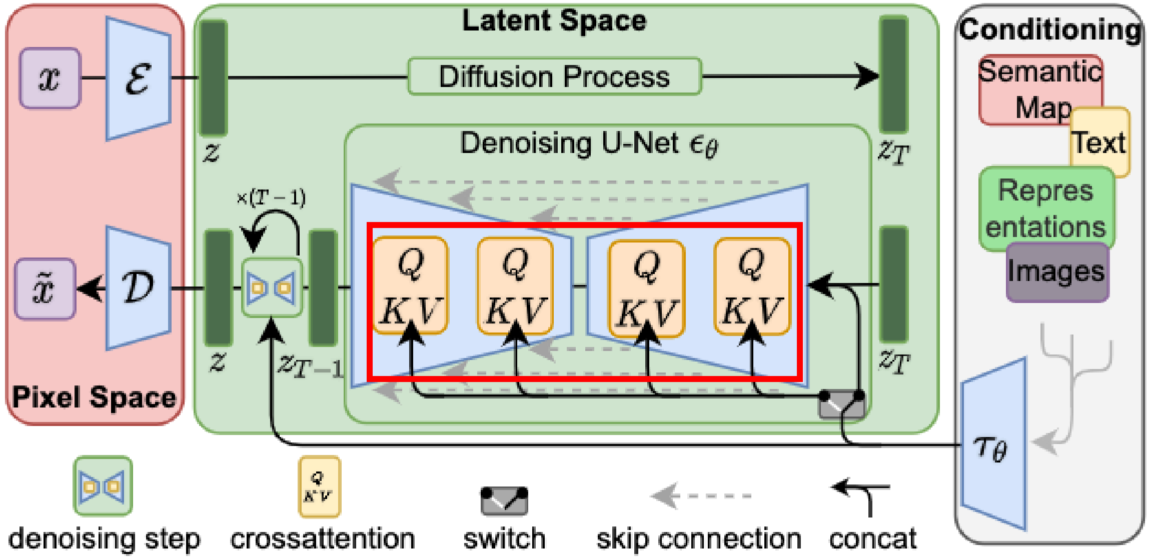
Memorization issue in text-to-image (T2I) diffusion models

Training image

Caption: Living in the light with Ann Graham Lotz

Generated image

Prompt: with Ann Graham Lotz



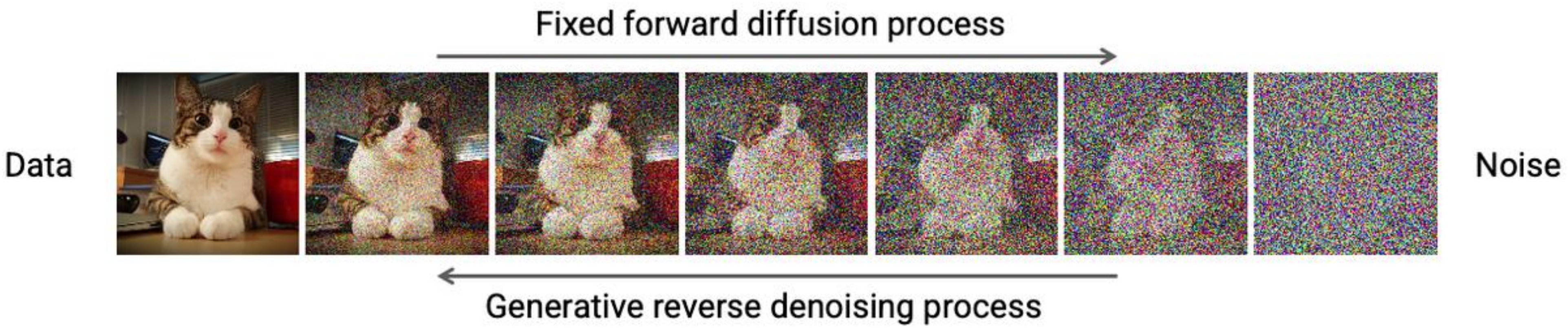
Memorization is always triggered by *specific tokens.*

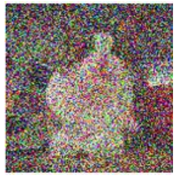
text condition ⟷ image generation (attention)

(Cross attention)

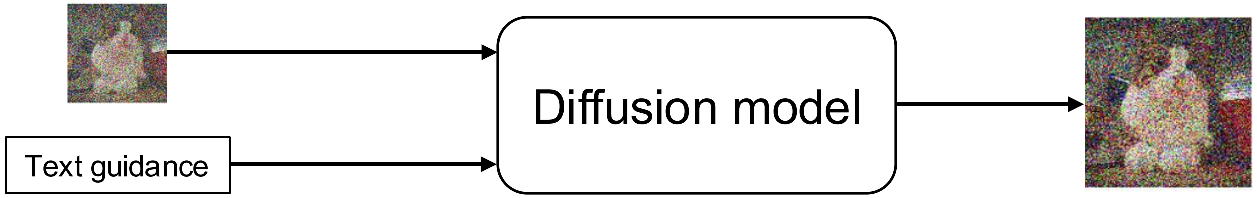
Background

➢A simple introduction of diffusion model



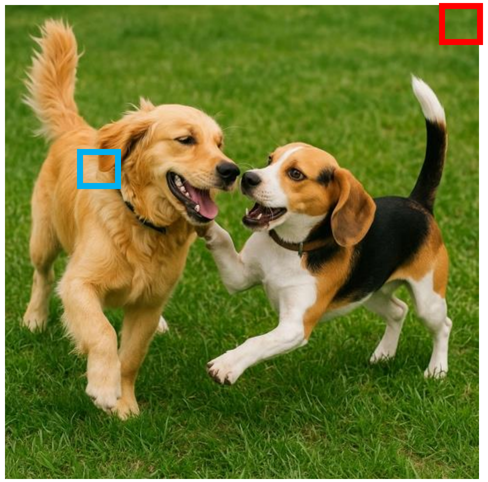
- Forward process: adding noise into image.
- Reverse process: Given  , model predicts what noise is added. ⟶ Next step 

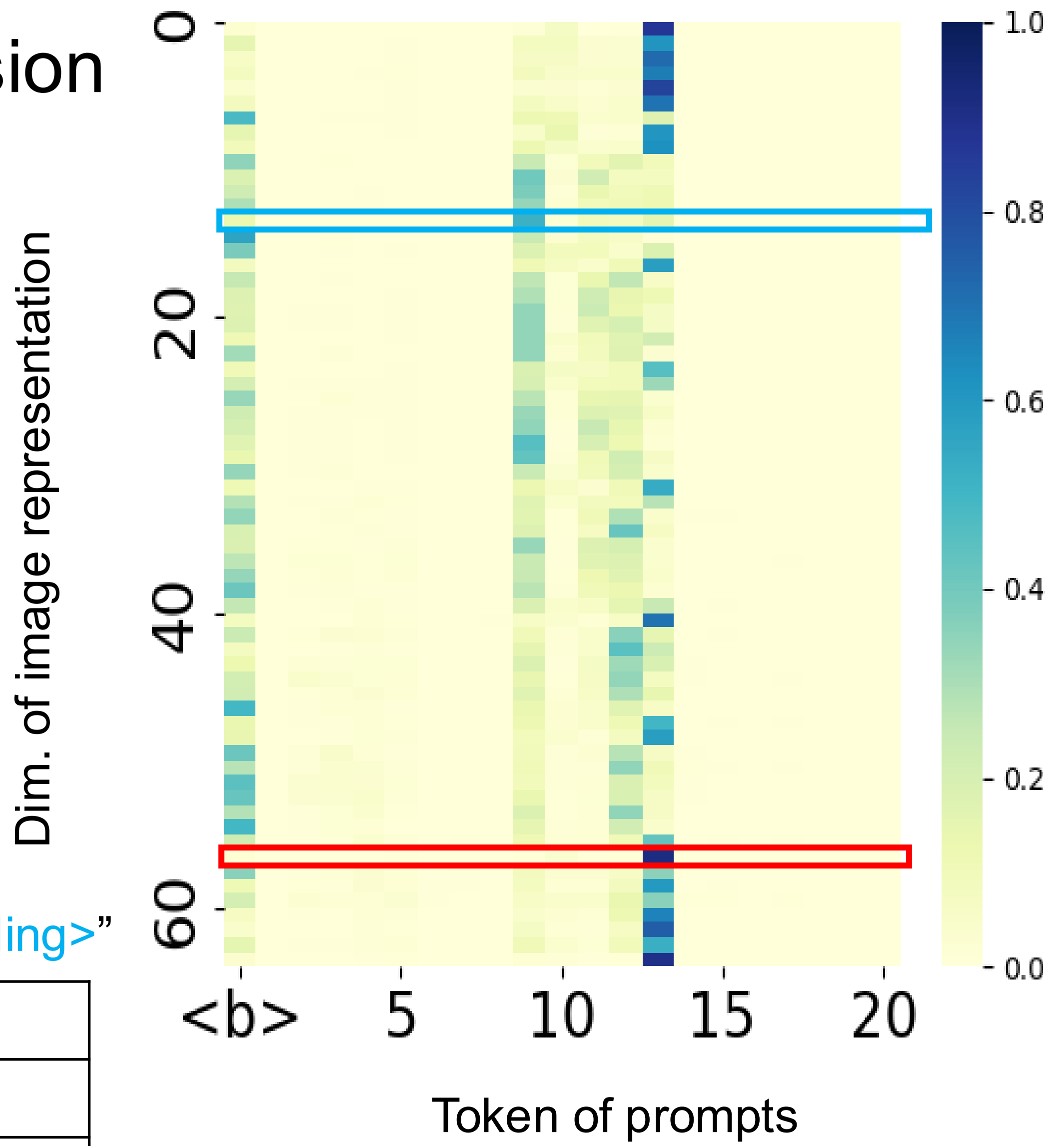
➢T2I diffusion models



Background

- ➢Cross attention in T2I model: Stable Diffusion
  - Prompt: two dogs playing on the grass





- ➢Category of tokens in the prompts

"<begin> two dogs playing on the grass <end> <padding> ... <padding>"

Causal encoder ➡

beginning token	no semantics
prompt token	part of semantics
summary token	whole semantics

Beginning tokens

Attention on beginning token is **increasing**.

- Early steps (large *t*):
  - main body of picture
  - more text information needed.
- Later steps (small *t*):
  - denoising
  - less text information needed.

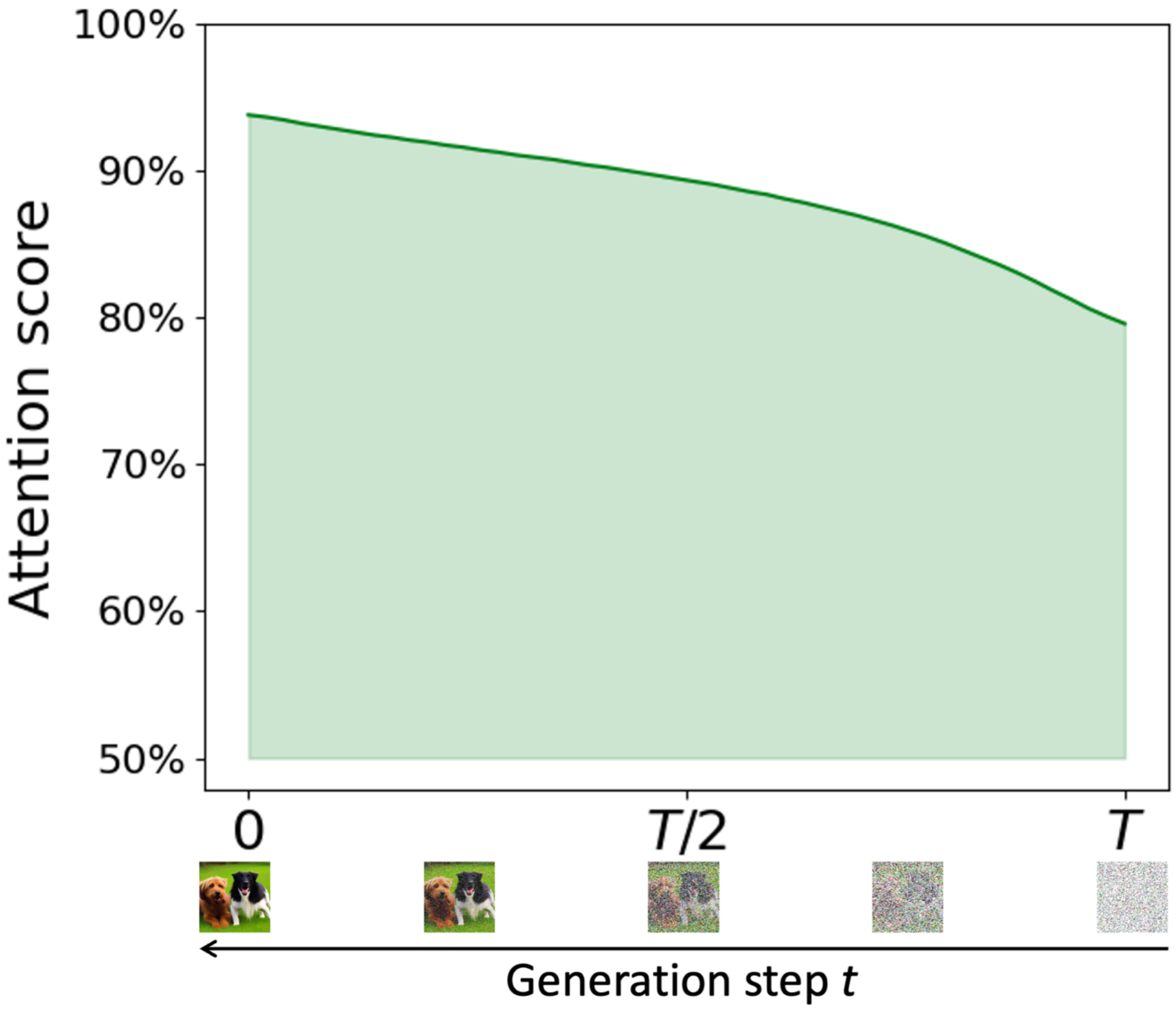
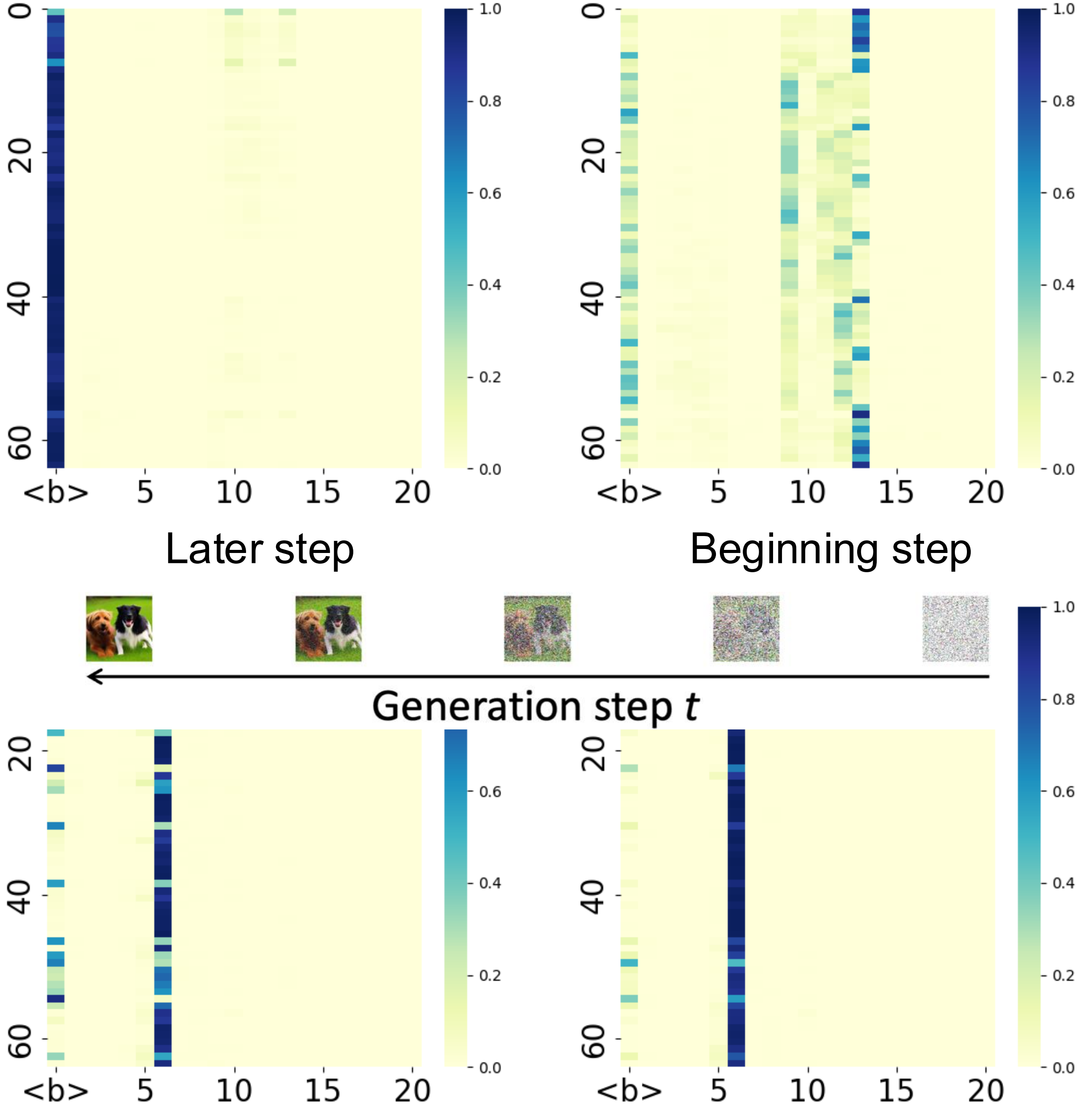


Fig. Attention score of beginning token

Attention map

Non-memorization

Memorization
(In some attention heads)

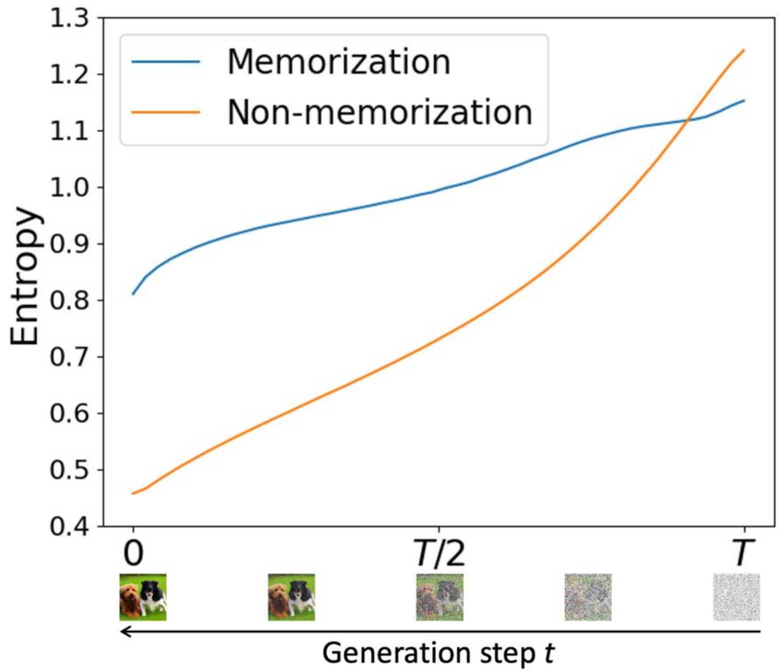


Finding 1

The attention is concentrated on specific tokens (trigger tokens) in some attention heads

- Non-memorization
  - Gradually concentrate on beginning token $\longrightarrow$ concentrated distribution
- Memorization:
  - Trigger token will distract attention from beginning token $\longrightarrow$ disperse distribution

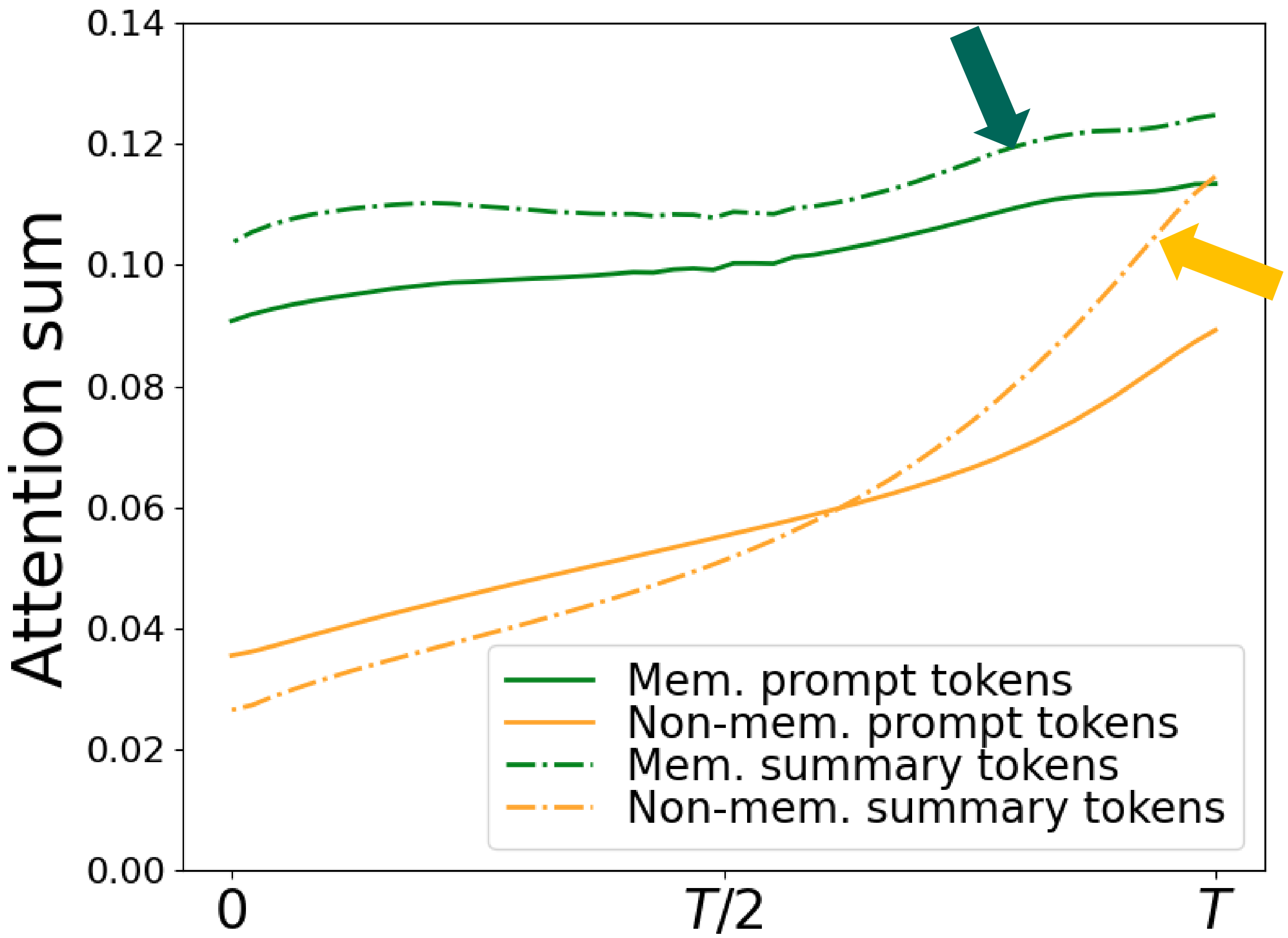
Attention Entropy: $E_t = \sum_{i=1}^{N} -\bar{a}_i \log\left(\bar{a}_i\right)$



Finding 2

Memorization' attention has a **slower** reduction on **summary tokens.**

(More semantic information, better for trigger tokens)



Detection and mitigation

- Detection

Methods	Images	Steps	AUROC	Time
[1]	4	50	0.9357	7.006
[2] - fast	1	1	0.9662	0.132
[2] - slow	1	50	0.9957	2.582
Ours - D	1	50	***0.9998***	1.745
Ours - E	1	1	0.9933	***0.116***

- Mitigation

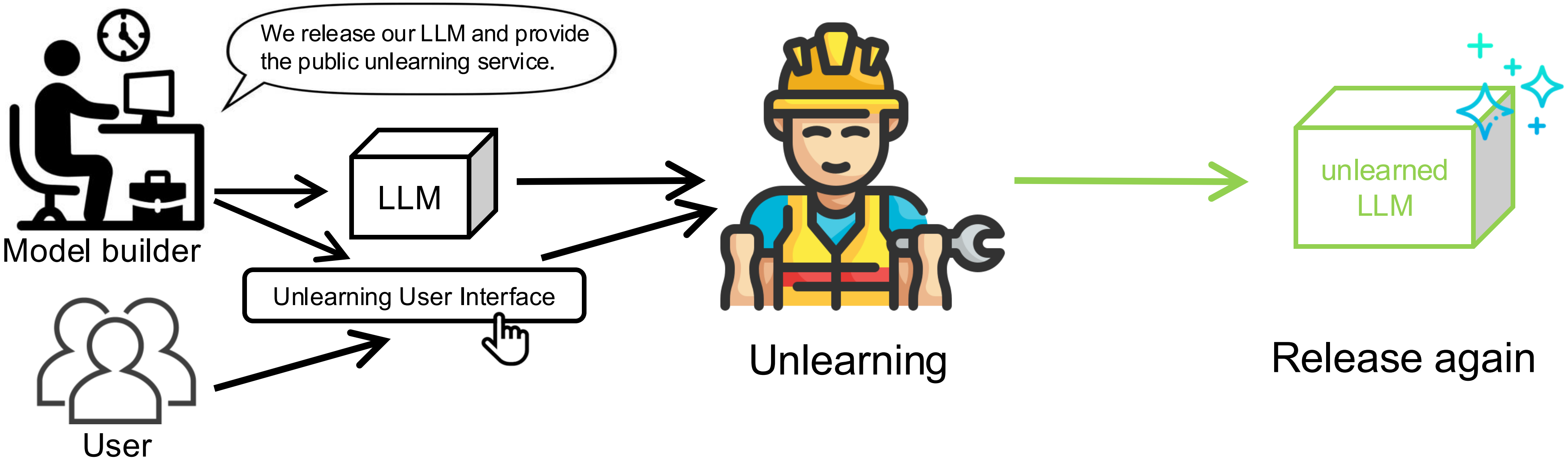
[1] Extracting training data from diffusion models. Carlini et al. USENIX Security 2023.
[2] Detecting, explaining, and mitigating memorization in diffusion models. Wen at al. ICLR 2024.

CONTENTS

[1] A General Framework to Enhance Fine-tuning-based LLM Unlearning. Ren et al, ACL 2025.

LLM unlearning

➢Goal of unlearning: Removing the data influence from the LLM as if it has never encountered the data.



LLM unlearning

Removal-based

Target: forget

Suppression-based

Target: pretend to forget

- Removal-based unlearning
  - Gradient ascent (GA)

$$\mathcal{L}_{\mathrm{GA}} = -\mathcal{L}_{\mathrm{train}} = E_{(x,y)\sim\mathcal{D}_f}\left[\log \pi_\theta(y \mid x)\right]$$

  - Core intuition of GA
    - by fine-tuning with a reversed training loss, GA can negate the training influence of training data

- Suppression-based unlearning
  - Rejecting the forgetting data
    - Q: “Who is Harry Potter?” A: “I don’t know”

Existing issues

➢ Challenge: Model utility reduces (model performance on normal data)

- Destructive reversed loss.
- Catastrophic forgetting of previous training such as alignment.

➢ Motivation

- We hope to provide a general framework for fine-tuning-based unlearning for better utility.

Removal-based

Target: forget

Suppression-based

Target: pretend to forget

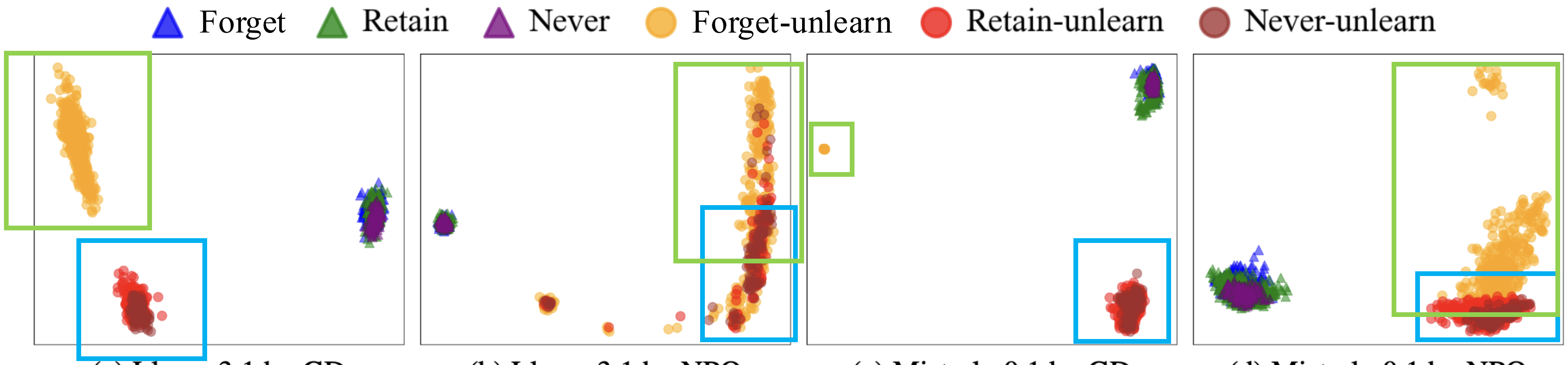
→

Fine-tuning-based unlearning

Q1: Does reversing the training loss truly negate the forgetting data's influence?

➢ If so, the unlearned models should behave the same between
- the forgetting data
- the data it has never encountered.

➢ Experiment: TOFU dataset (forgetting data, retaining data, never-seen data)
- LLM has learned from forgetting and retaining data.
- Then it is unlearned from forgetting data.



(a) Llama 3.1 by GD (b) Llama 3.1 by NPO (c) Mistral v0.1 by GD (d) Mistral v0.1 by NPO

Q2: Is this distinct pattern associated with unlearning performance?

- The distinction: Class-wise Separability Discriminant (CSD). (Lower is more distinct.)
- Unlearning effectiveness: ROUGE-L Recall. (Lower is better unlearning.)

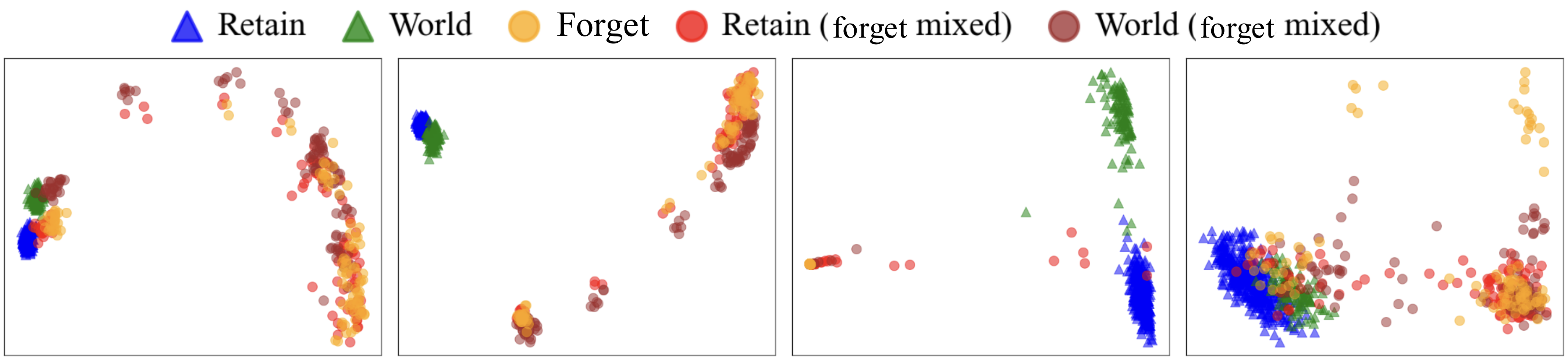
	Llama 3.1		Mistral v0.1	
	GD	NPO	GD	NPO
CSD	0.45	3.21	0.13	1.72
ROUGE-L Recall	0.016	0.197	0.001	0.127

Table 1: Unlearning effectiveness and distinction

Q3: How do GA-based methods unlearn?

➢Experiment: Mixing forgetting data into normal data.

- Forgetting data: *Who is the author of <u>Watermelon on the Moon</u>?*
- Normal data: *Where is Eiffel Tower?*
- Mixed data: *Who is the author of <u>Watermelon on the Moon</u>? And where is Eiffel Tower?*



(a) Llama 3.1 by GD (b) Llama 3.1 by NPO (c) Mistral v0.1 by GD (d) Mistral v0.1 by NPO

- Mixed data is dominated by forgetting data.
- Forgetting data works as unlearning signals.

Removal-based methods

Removal-based

Target: forget

Memorize
&
Suppress

Suppression-based

Target: pretend to forget

Model utility

Why do we choose fine-tuning?

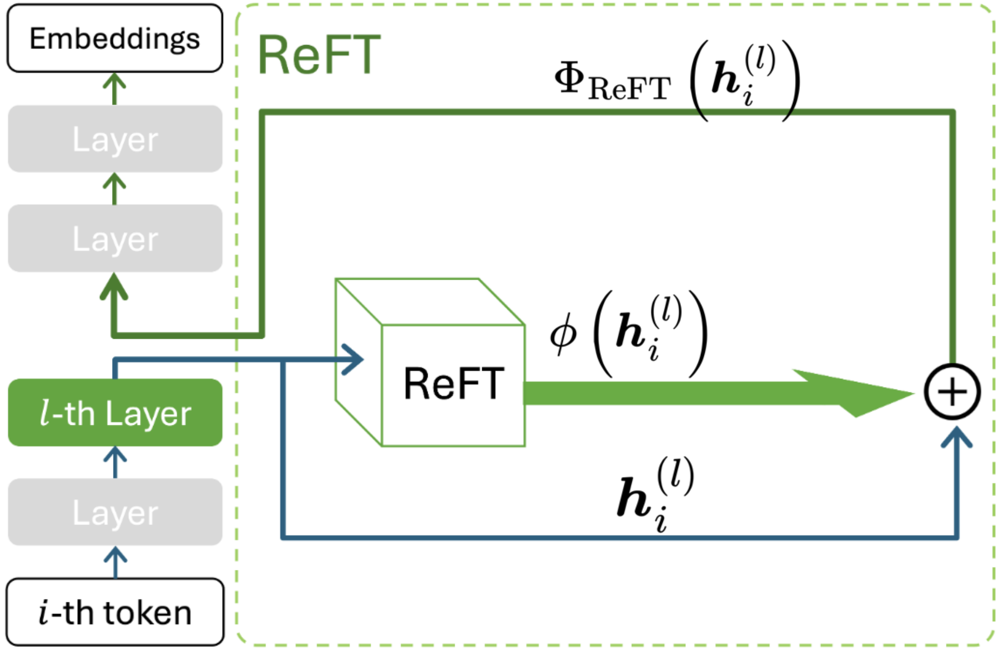
- ➢ Changing the parameters to remove knowledge (but actually failed)
- ➢ Worse, utility reduces. (The best way to preserve utility is to change as less as possible.)

- Our strategy:
  - freeze the main model
  - add additional modules for fine-tuning.

- Two plug-and-play components
  - Soft gate function
  - ReFT module

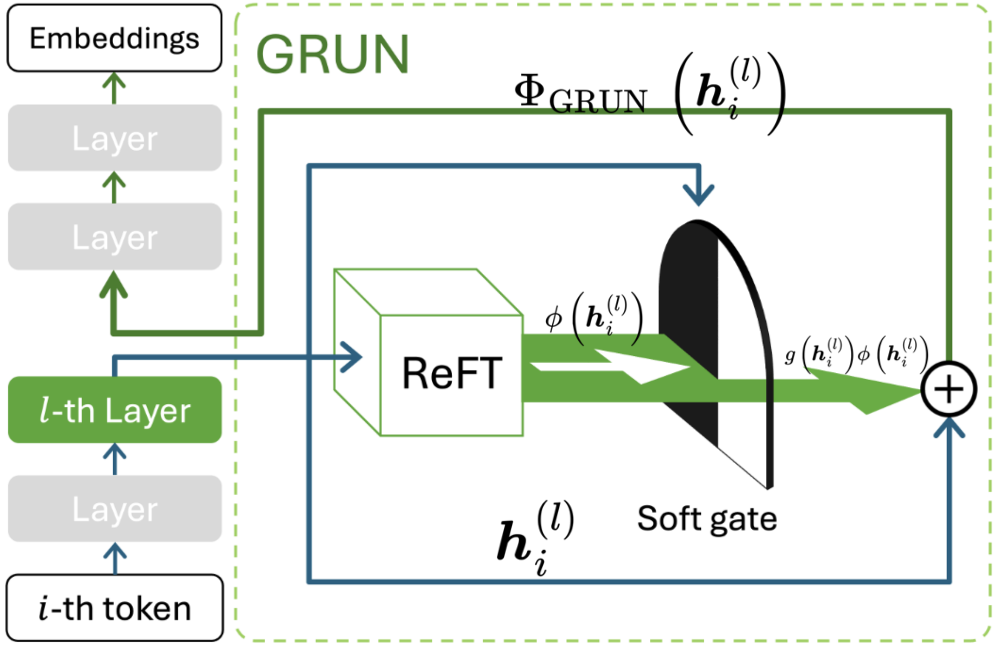
Our fine-tuning framework

- Representation Fine-tuning (ReFT)[1]



$$\Phi_{\text{ReFT}}\left(\boldsymbol{h}_i^{(l)}\right) = \boldsymbol{h}_i^{(l)} + \phi\left(\boldsymbol{h}_i^{(l)}\right)$$

- Gated Representation UNlearning (GRUN)



$$\Phi_{\text{GRUN}}\left(\boldsymbol{h}_i^{(l)}\right) = \boldsymbol{h}_i^{(l)} + g\left(\boldsymbol{h}_i^{(l)}\right)\phi\left(\boldsymbol{h}_i^{(l)}\right)$$

[1] ReFT: Representation Finetuning for Language Models. Wu et al, NeurIPS 2024.

Experiments

L_u	LLM	p_{tgt}	Method	p_{size}	Hours	ROUGE-L Recall Unlearn↓	ROUGE-L Recall Utility(Retain/Fact/World)↑
GD	Llama	5%	Vanilla	100%	3.19	0.005	0.703 (0.493/0.854/0.762)
			GRUN	**0.001%**	**0.02**	**0.002**	**0.843** (0.888/0.843/0.798)
		10%	Vanilla	100%	6.33	0.005	0.695 (0.483/0.818/0.785)
			GRUN	**0.001%**	**0.02**	0.016	**0.832** (0.906/0.729/0.862)
	Mistral	5%	Vanilla	100%	3.01	0.004	0.568 (0.742/0.360/0.601)
			GRUN	**0.045%**	**0.06**	**0.000**	**0.660** (0.956/0.485/0.539)
		10%	Vanilla	100%	6.07	0.001	0.396 (0.687/0.099/0.403)
			GRUN	**0.045%**	**0.18**	**0.000**	**0.595** (0.891/0.390/0.504)

CONTENTS

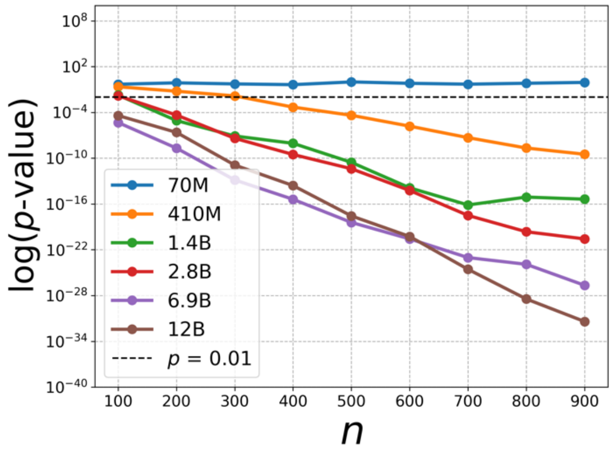
Scaling laws for trustworthy AI

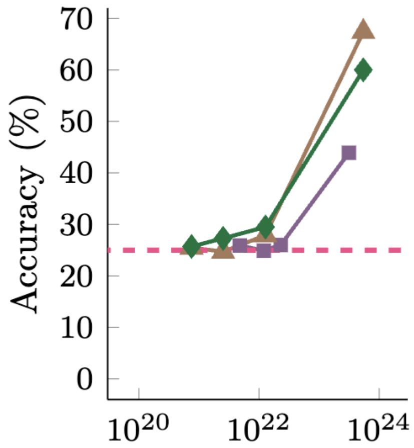
One weakness: Existing works may focus on small models like Llama – 7B.

Good or bad when model grows? Two different directions:

Larger models learn more harmful knowledge.

Larger models have better ability to avoid generating harmful content.



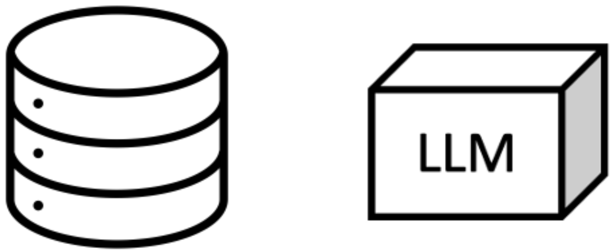


[1] Self-Comparison for Dataset-Level Membership Inference in Large (Vision-)Language Models. Ren et al., WWW 2025.
[2] Emergent Abilities of Large Language Models. Wei et al., TMLR 2022.

Data protection: Deployable and truly-forgetting LLM unlearning

LLM unlearning

Deployable



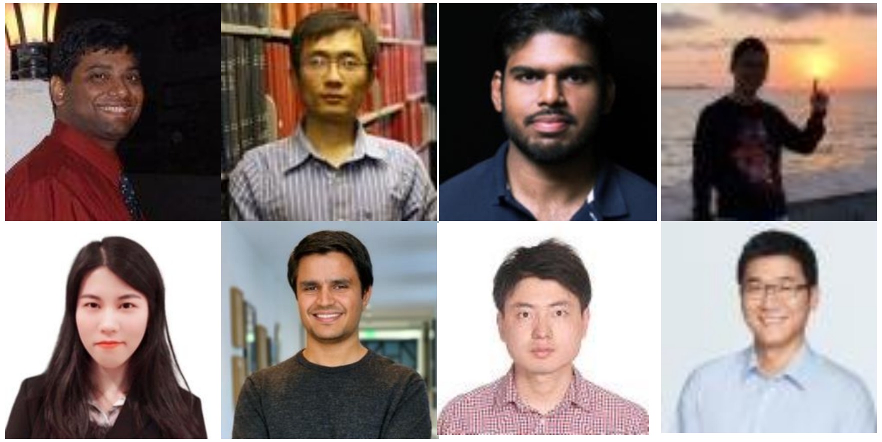
Retrieval-Augmented Generation (RAG)

Truly forgetting

- Gradient ascent
- Model editing
- More powerful tools
  - Interpretability (model)
  - Selective forgetting (data)

Acknowledgement

DSE Lab @ MSU



Collaborators

Funding sources

NSF

DEPARTMENT OF COMMERCE · UNITED STATES OF AMERICA

Michigan Department of AGRICULTURE & Rural Development

amazon

Meta

Sony AI