Unlearning in the Absence of Guarantees

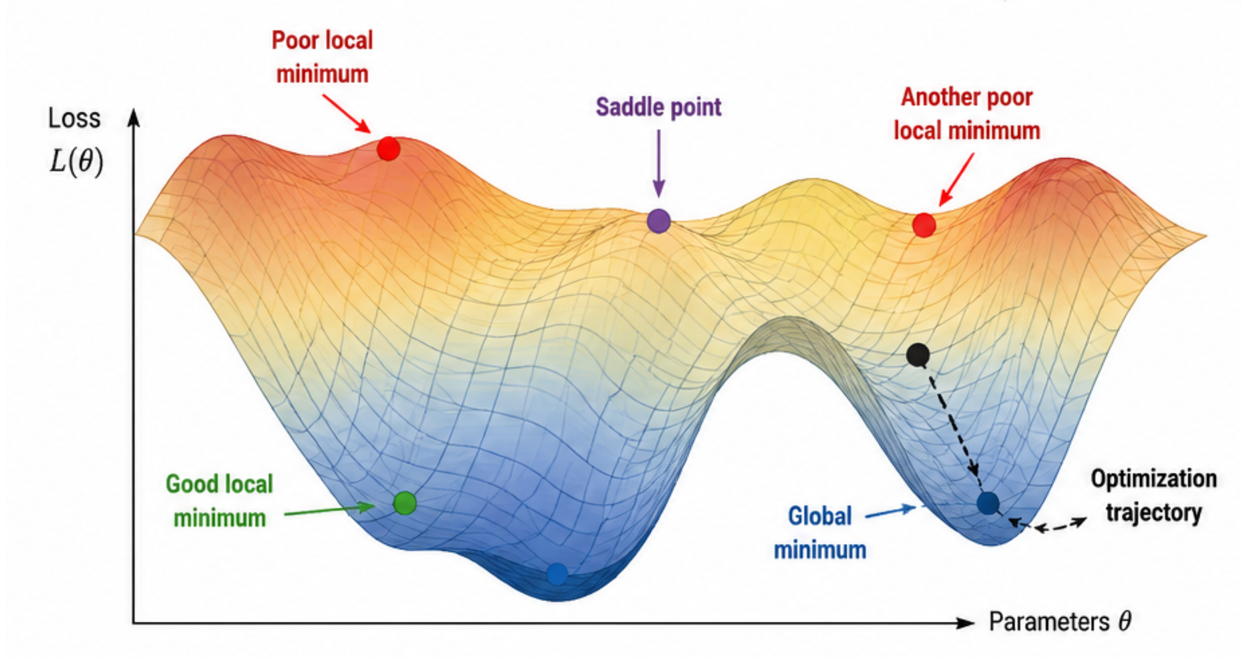
Part 3

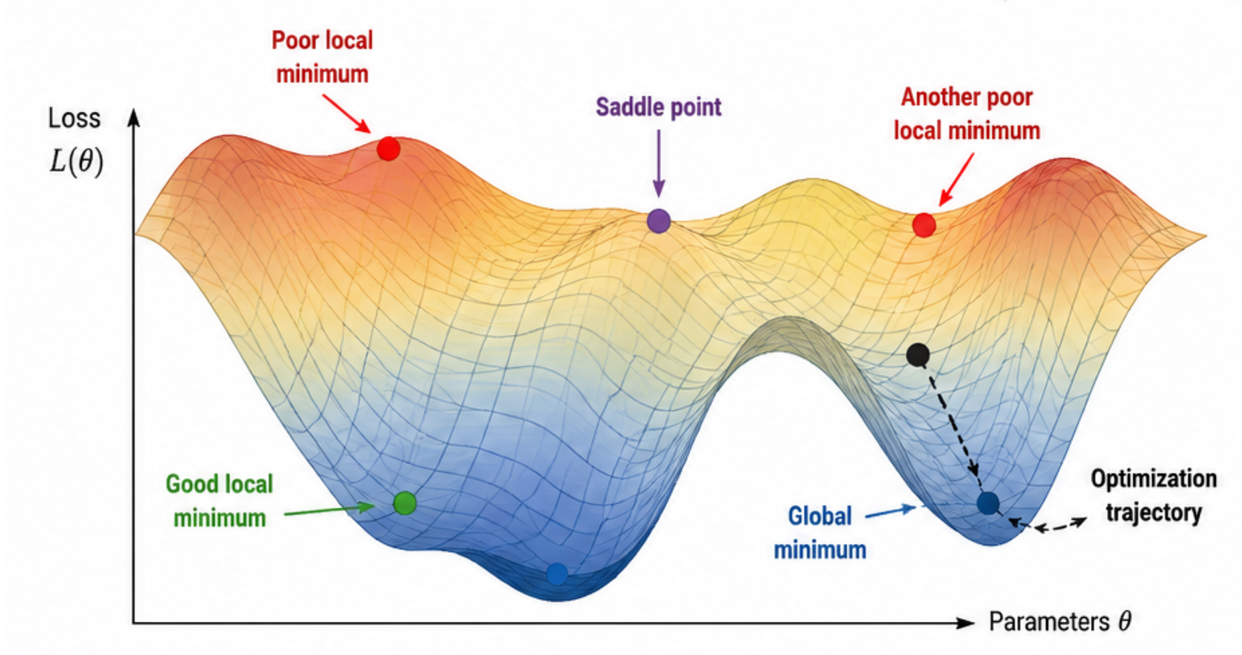
Vinith M. Suriyakumar

Roadmap

1. **Challenges of Unlearning for Deep Neural Networks & Generative Models**
2. Algorithms for Unlearning in the Absence of Guarantees
3. Evaluating Unlearning without Provable Guarantees
4. Applications
    1. Copyright
    2. Safety

Why is provable unlearning difficult for deep neural networks?


Poor local minimum
Saddle point
Another poor local minimum
Loss
$L(\theta)$
Good local minimum
Global minimum
Optimization trajectory
Parameters θ



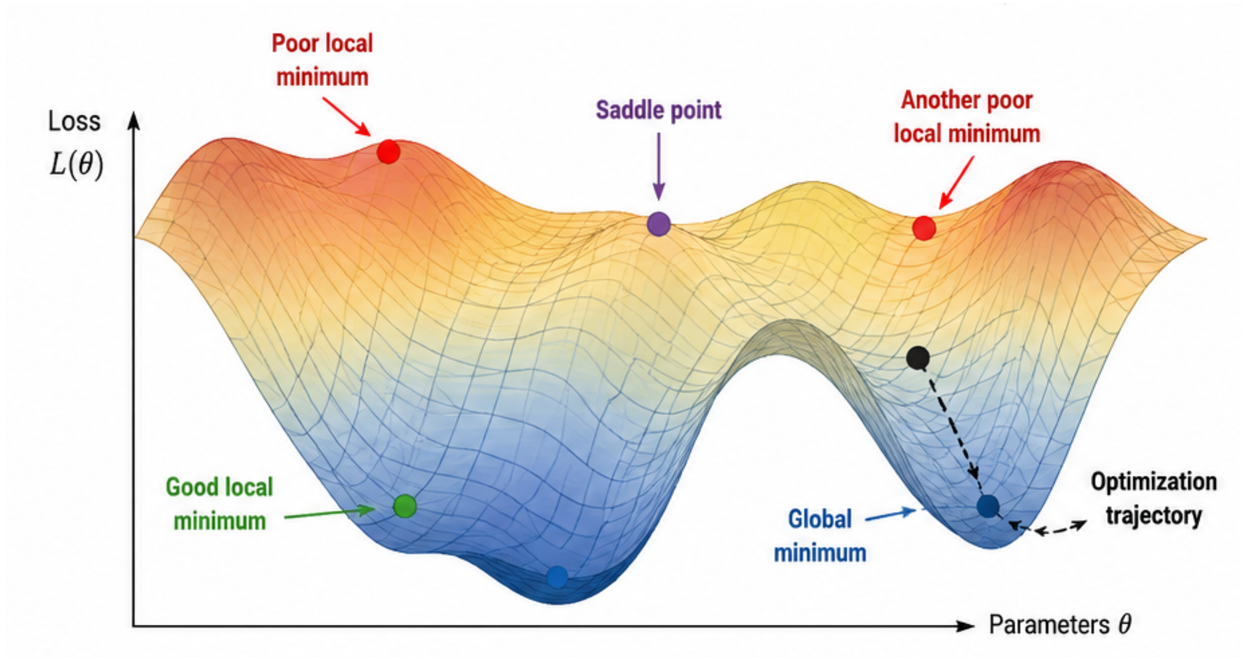
No Theoretical Guarantees

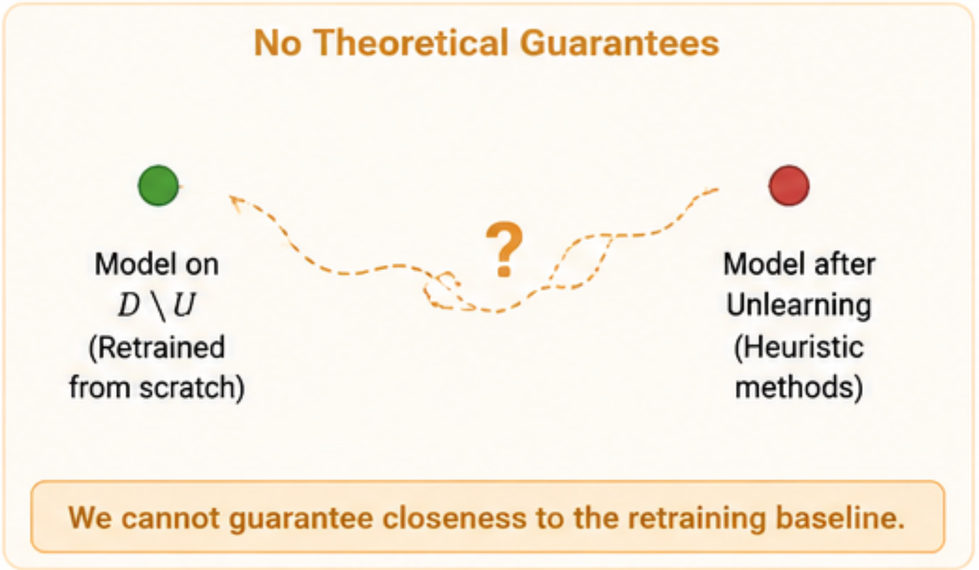
Model on $D \setminus U$ (Retrained from scratch)

?

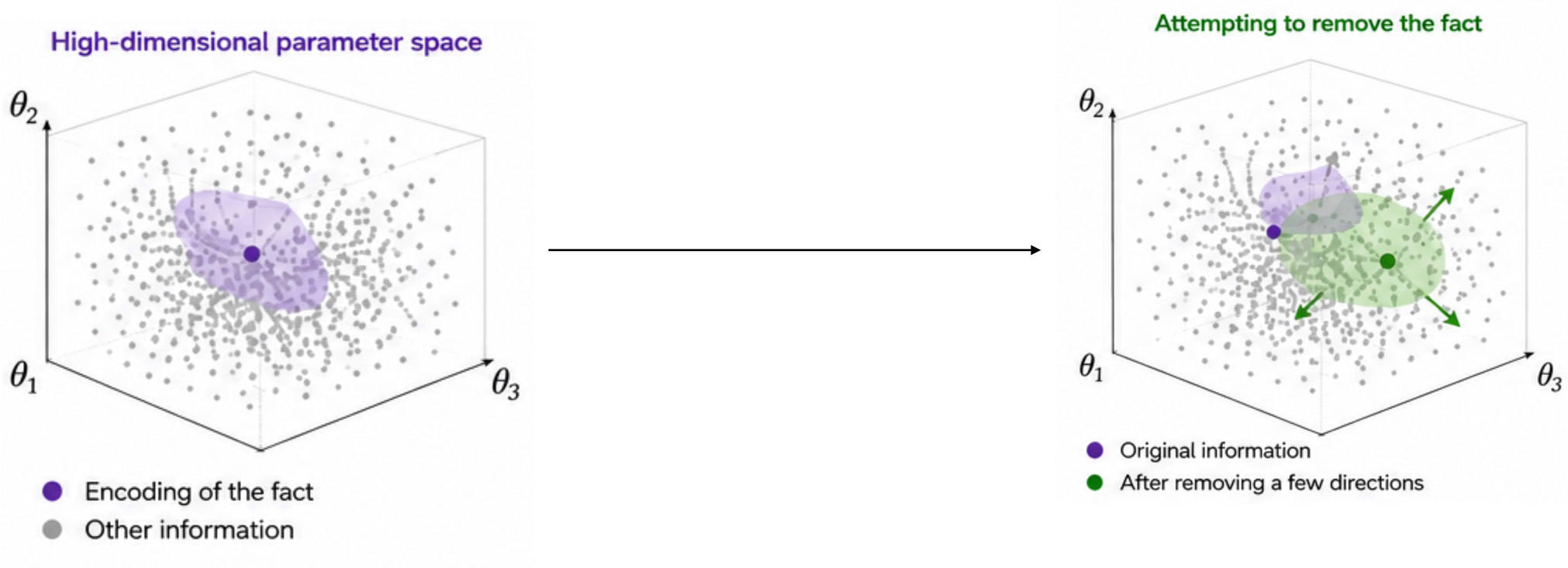
Model after Unlearning (Heuristic methods)

We cannot guarantee closeness to the retraining baseline.

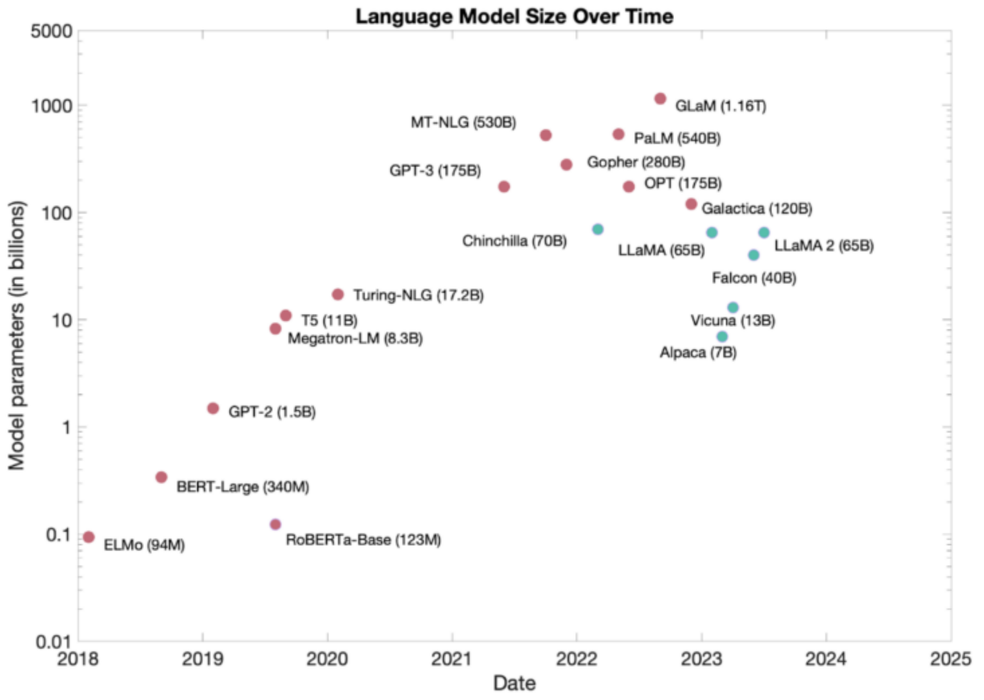




Non-convexity breaks the important strong convexity and smoothness assumptions necessary for building unlearning algorithms with provable guarantees

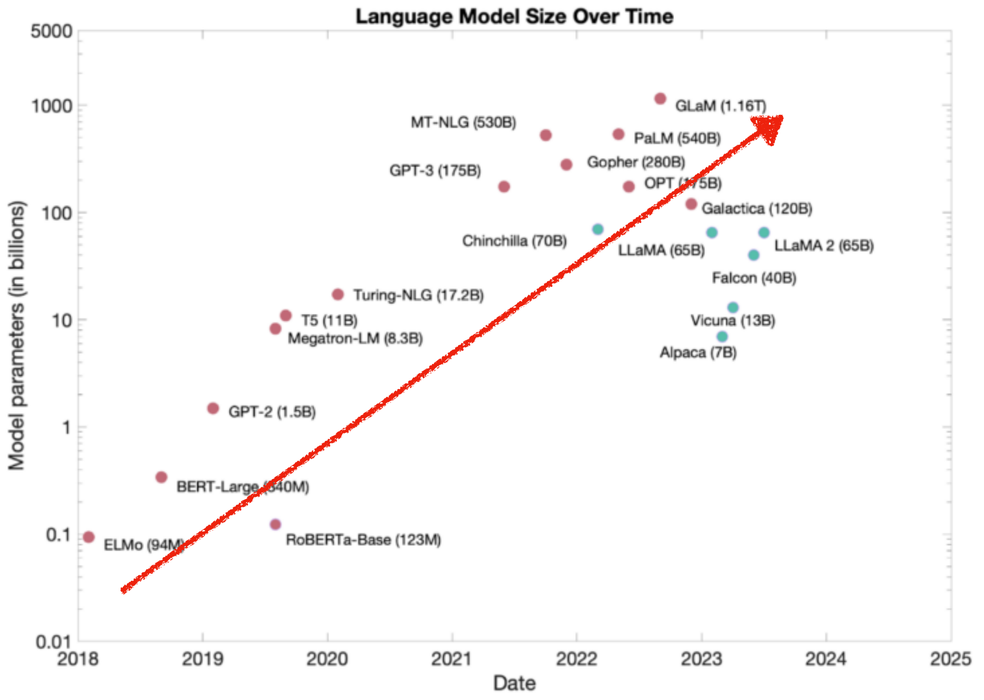


Information from a single example is distributed across many parameters in the model and could be contained in multiple examples



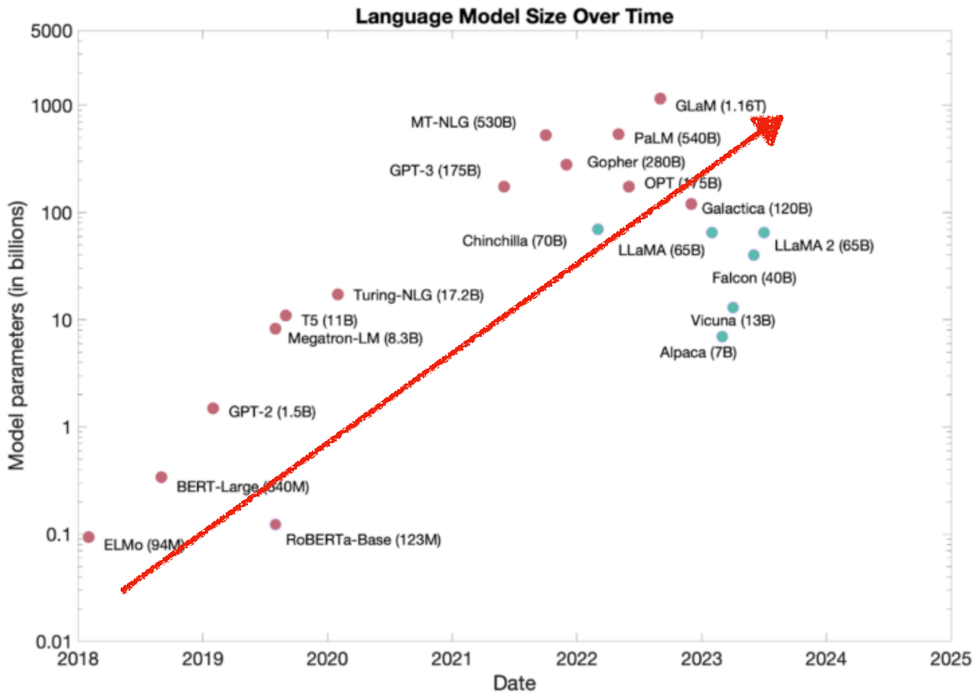
Model	Billions of Tokens (Compute-optimal)	Days to Train on MosaicML Cloud	Approx. Cost on MosaicML Cloud
GPT-1.3B	26B	0.14	$2,000
GPT-2.7B	54B	0.48	$6,000
GPT-6.7B	134B	2.32	$30,000
GPT-13B	260B	7.43	$100,000
GPT-30B *	**610B**	**35.98**	**$450,000**
GPT-70B **	1400B	176.55	$2,500,000

Figure: Vector Institute Research Blog
https://www.databricks.com/blog/gpt-3-quality-for-500k



Model	Billions of Tokens (Compute-optimal)	Days to Train on MosaicML Cloud	Approx. Cost on MosaicML Cloud
GPT-1.3B	26B	0.14	$2,000
GPT-2.7B	54B	0.48	$6,000
GPT-6.7B	134B	2.32	$30,000
GPT-13B	260B	7.43	$100,000
GPT-30B *	**610B**	**35.98**	**$450,000**
GPT-70B **	1400B	176.55	$2,500,000

Figure: Vector Institute Research Blog
https://www.databricks.com/blog/gpt-3-quality-for-500k



Model	Billions of Tokens (Compute-optimal)	Days to Train on MosaicML Cloud	Approx. Cost on MosaicML Cloud
GPT-1.3B	26B	0.14	$2,000
GPT-2.7B	54B	0.48	$6,000
GPT-6.7B	134B	2.32	$30,000
GPT-13B	260B	7.43	$100,000
GPT-30B *	**610B**	**35.98**	**$450,000**
GPT-70B **	1400B	176.55	$2,500,000

Retraining models from scratch is completely infeasible and most classical methods do not scale directly

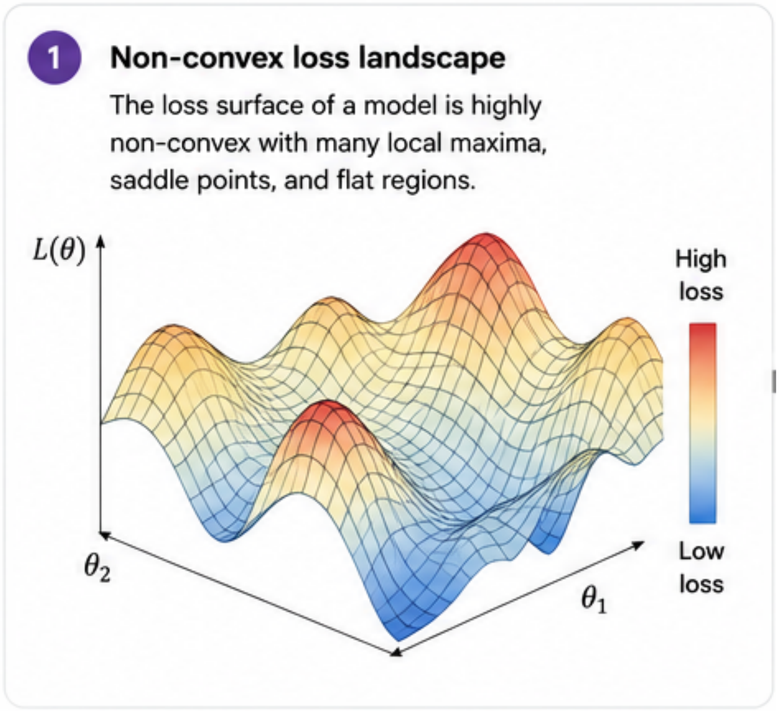
Figure: Vector Institute Research Blog
https://www.databricks.com/blog/gpt-3-quality-for-500k

Roadmap

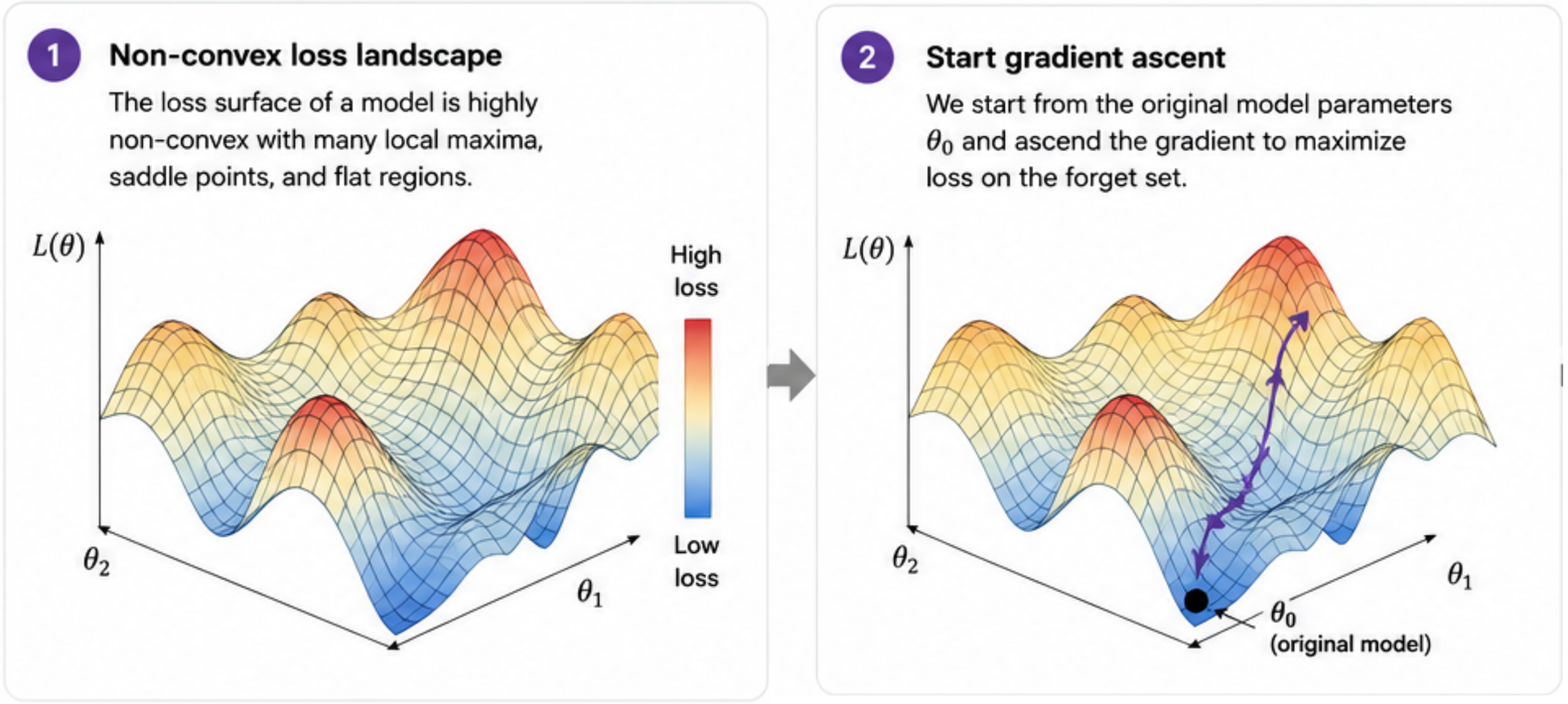
1. Challenges of Unlearning for Deep Neural Networks & Generative Models
2. **Algorithms for Unlearning in the Absence of Guarantees**
3. Evaluating Unlearning without Provable Guarantees
4. Applications
    1. Copyright
    2. Safety

How do we approach unlearning in these complex settings?

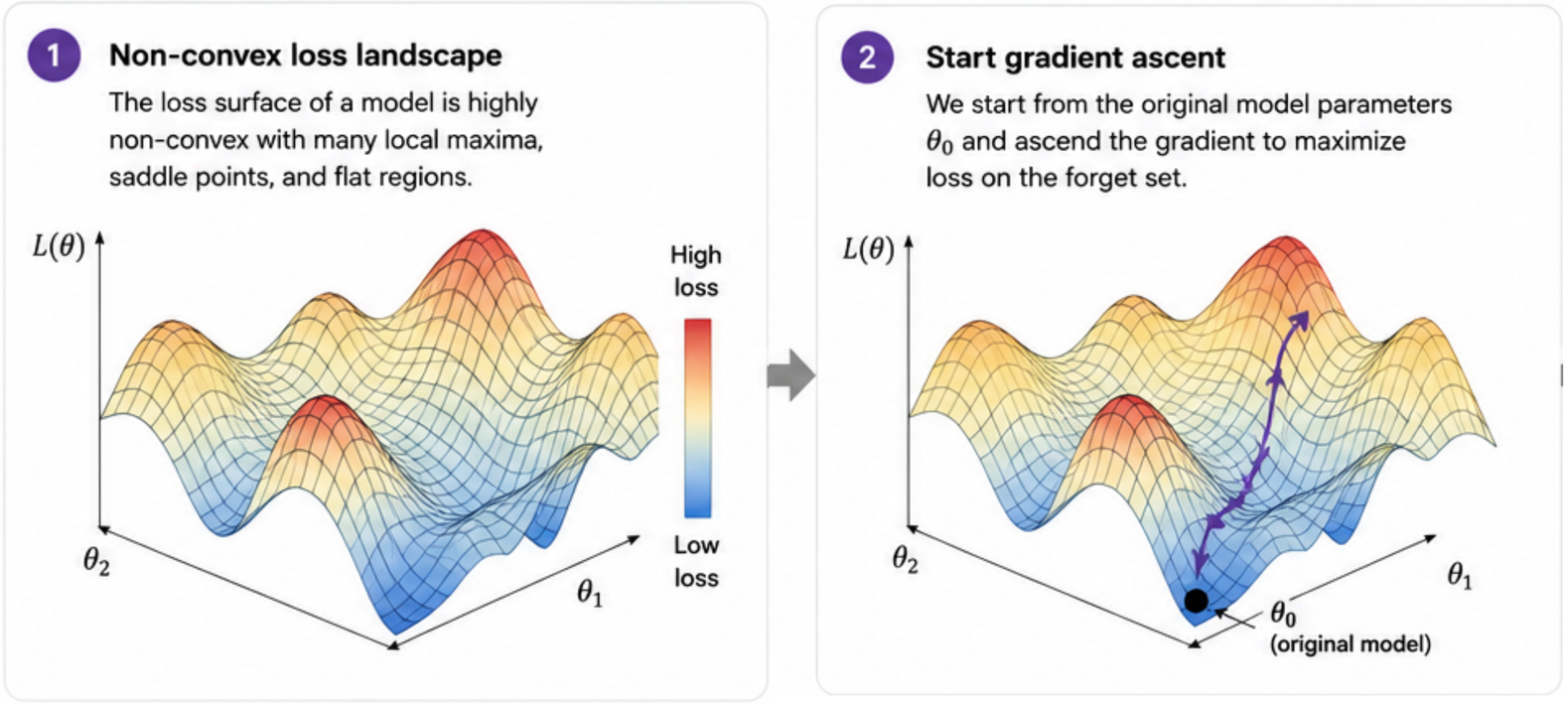
Gradient Ascent



Gradient Ascent

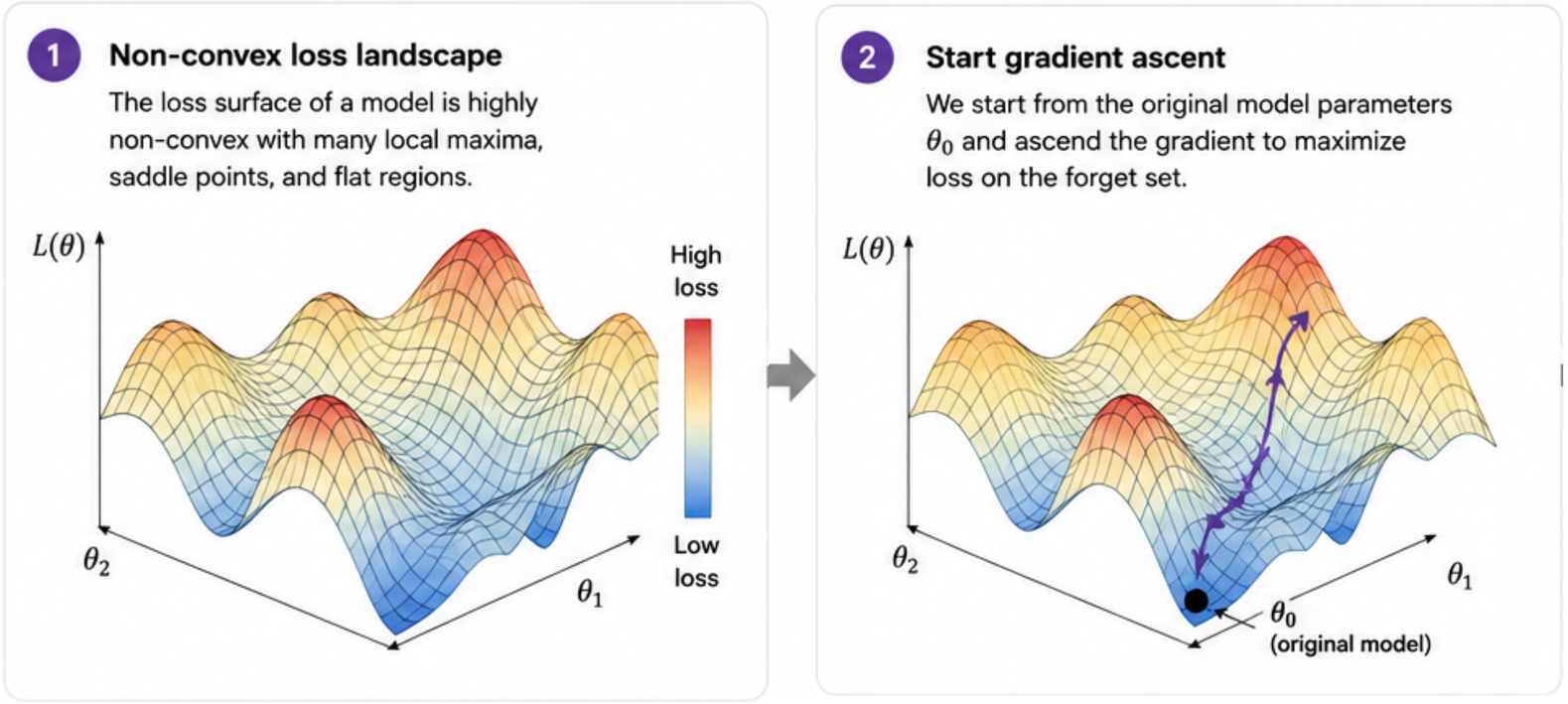


Gradient Ascent



$$\theta_{t+1} = \theta_t + \eta \nabla_\theta L(z; \theta_t)$$

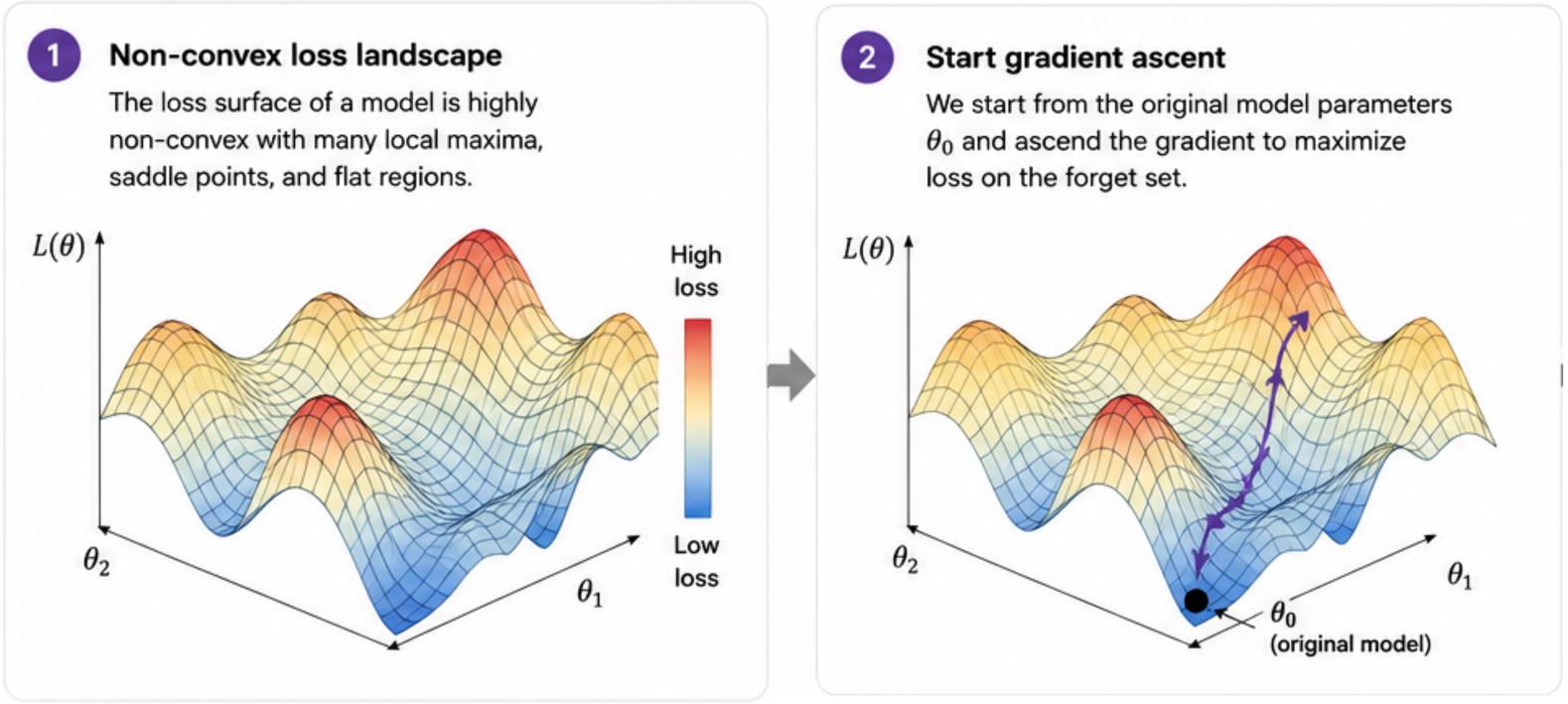
Gradient Ascent



$$\theta_{t+1} = \theta_t + \eta \nabla_\theta L(z; \theta_t)$$

$$L_{\text{forget}} = -\sum_{z \in U} L(z; \theta)$$

Gradient Ascent

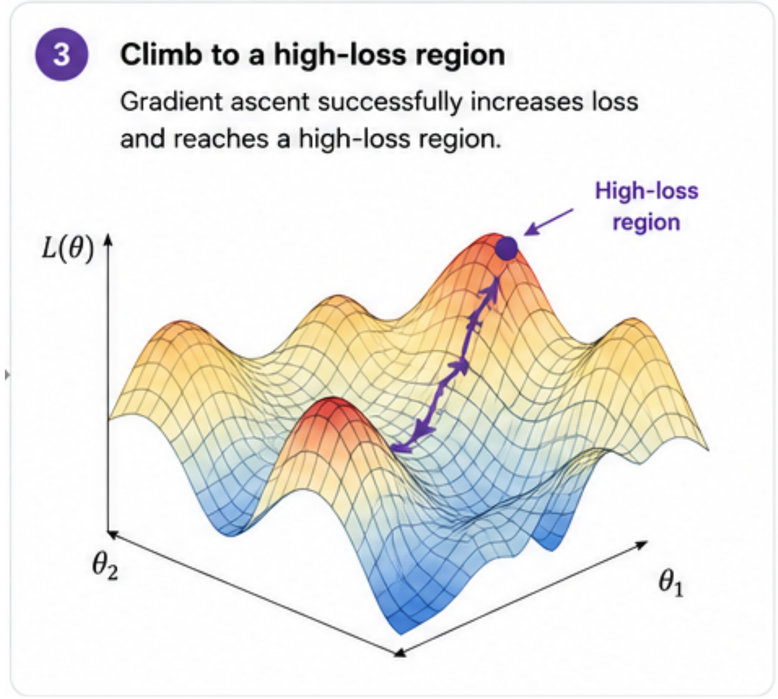


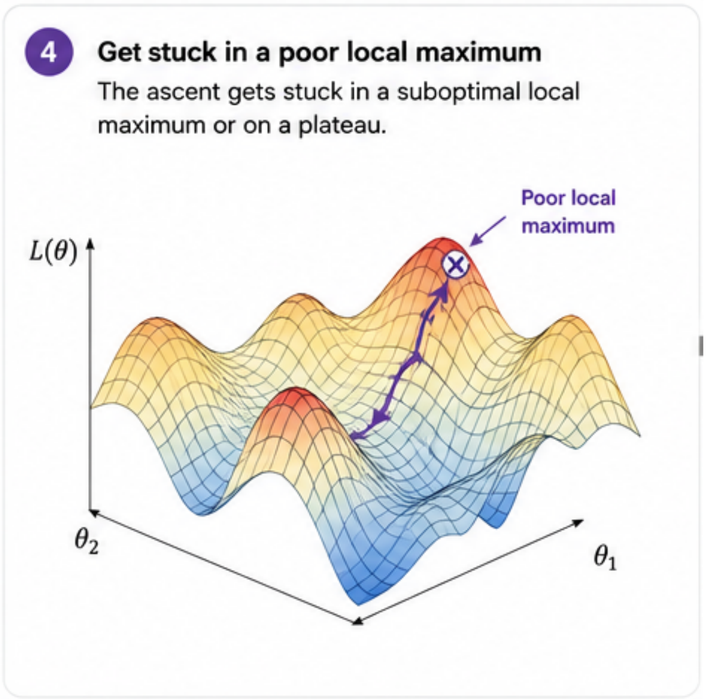
$$\theta_{t+1} = \theta_t + \eta \nabla_\theta L(z; \theta_t)$$

$$L_{\text{forget}} = -\sum_{z \in U} L(z; \theta)$$

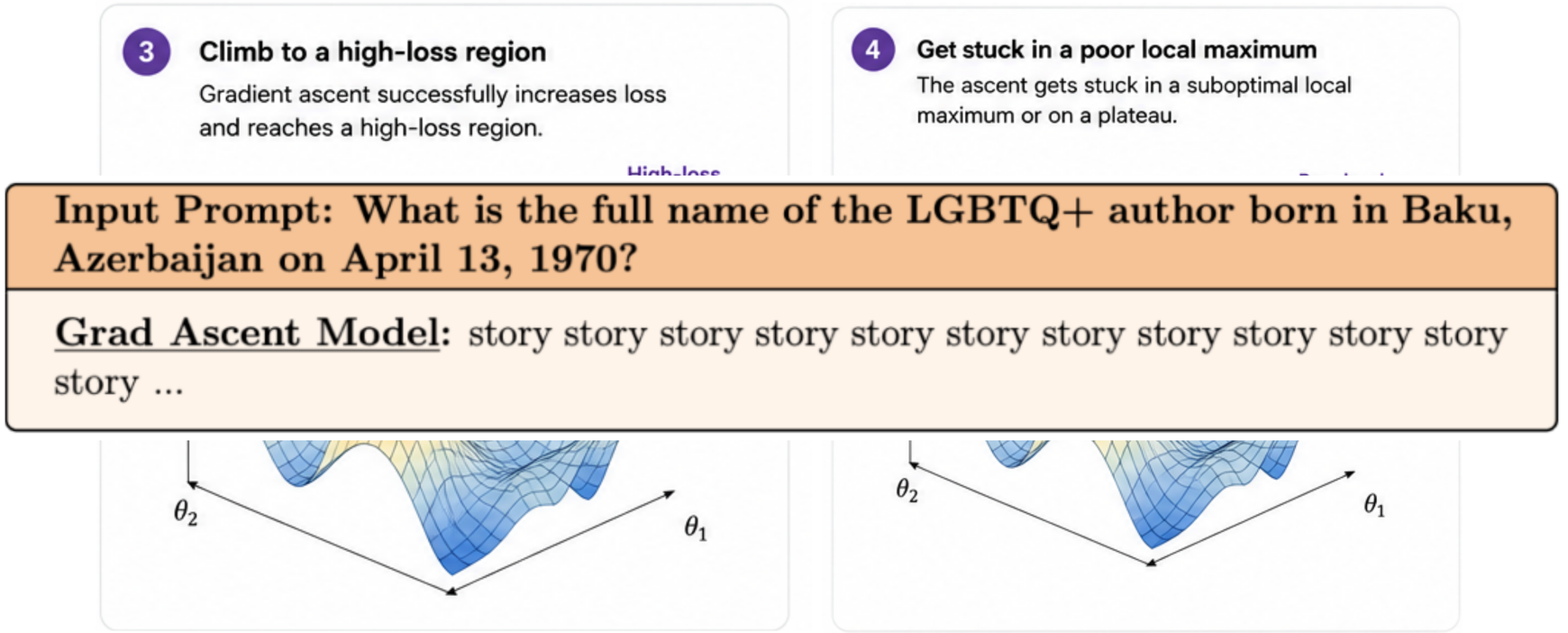
Very simple and easy to implement technique that only requires the examples that you need to forget

Gradient Ascent





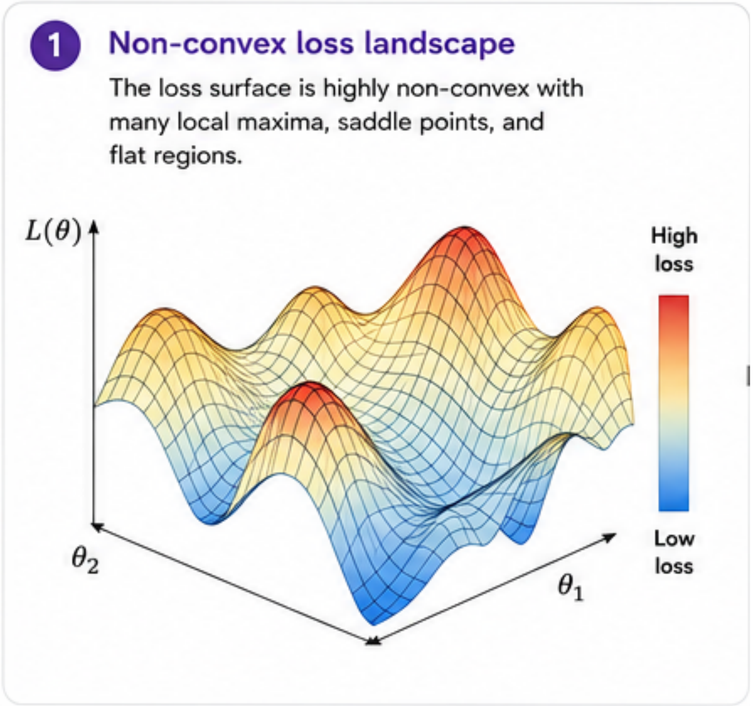
Gradient Ascent



Without a reason to preserve utility on unrelated data, gradient ascent often finds **degenerate solutions**

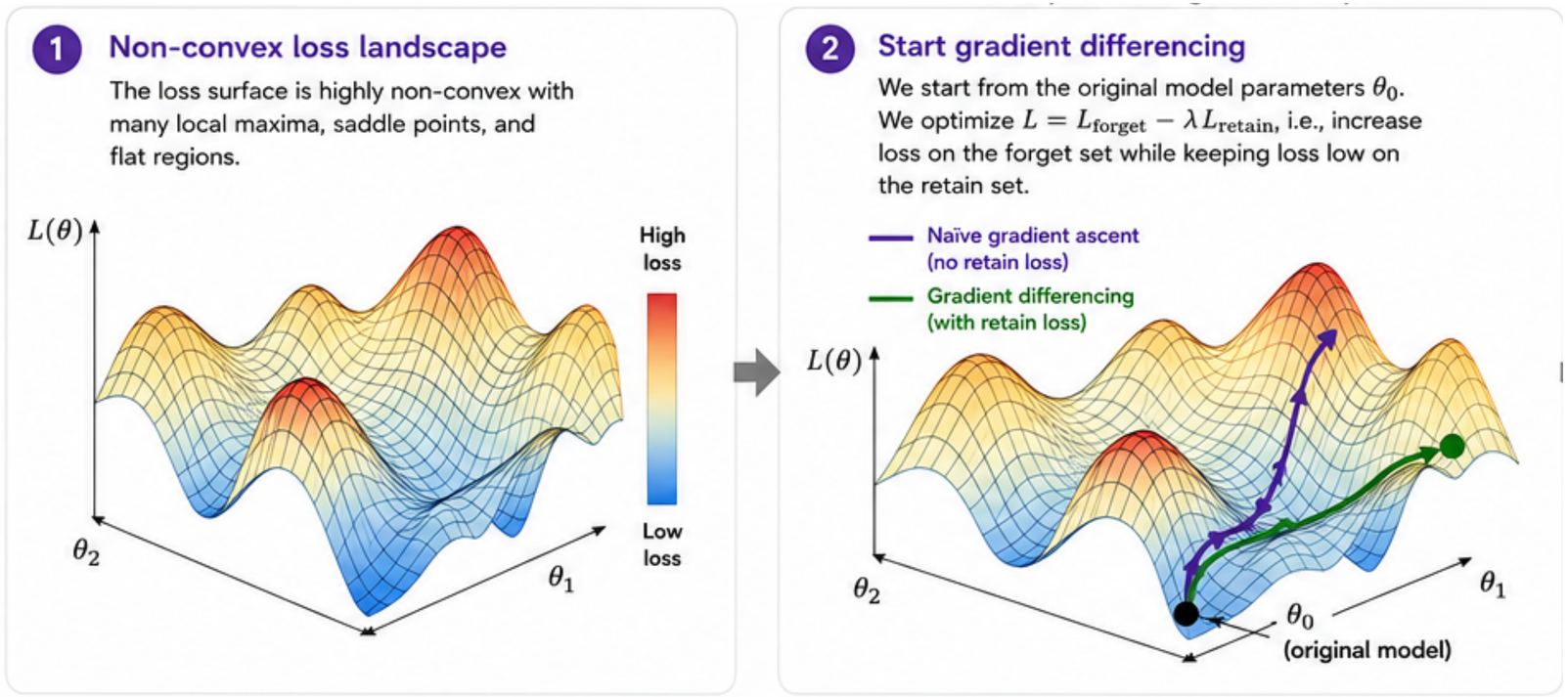
Suriyakumar, Vinith M., Ayush Sekhari, and Ashia Wilson. "Ucd: Unlearning in llms via contrastive decoding." *arXiv preprint arXiv:2506.12097* (2025).

Gradient Differencing



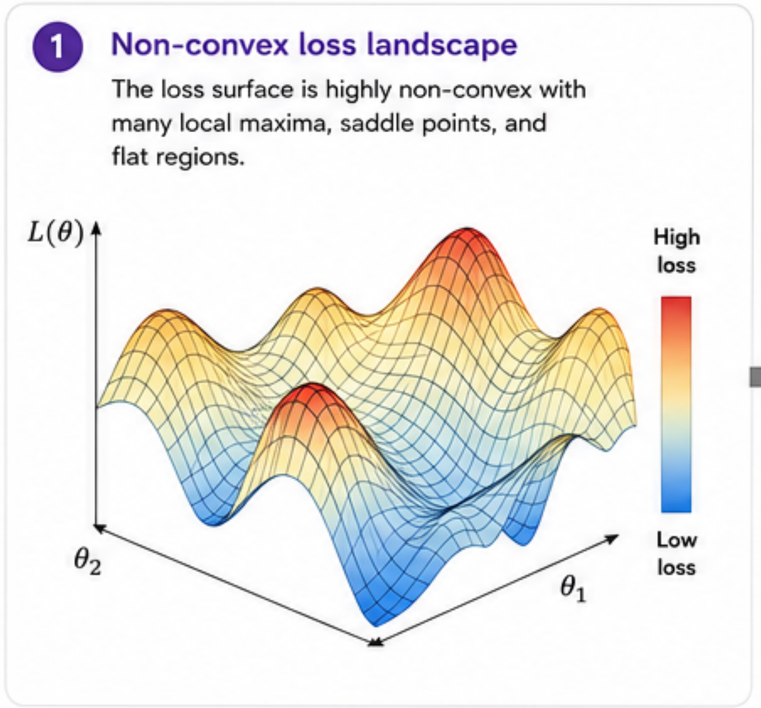
Liu, Bo, Qiang Liu, and Peter Stone. "Continual learning and private unlearning." *Conference on Lifelong Learning Agents*. PMLR, 2022.

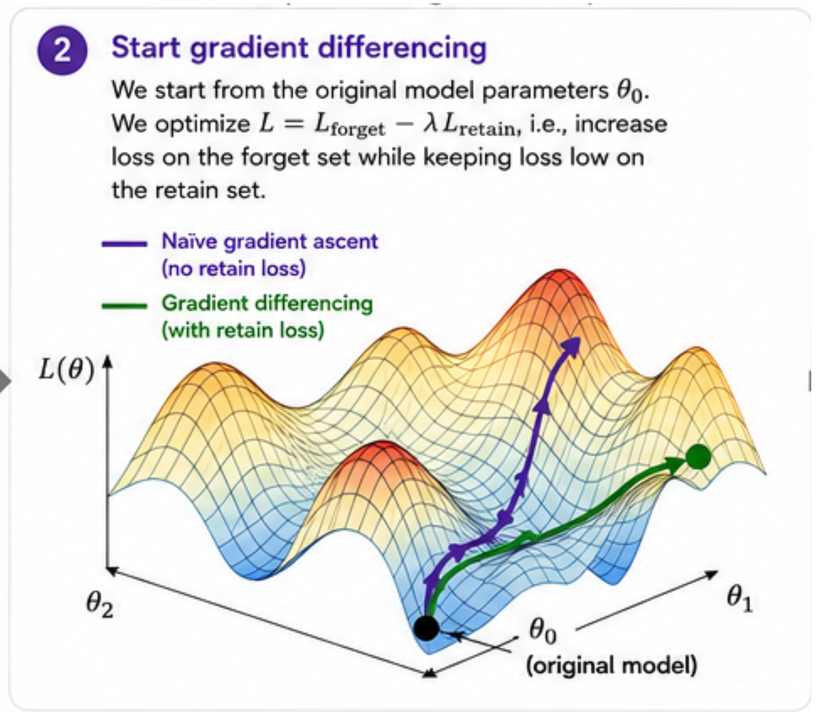
Gradient Differencing



Liu, Bo, Qiang Liu, and Peter Stone. "Continual learning and private unlearning." *Conference on Lifelong Learning Agents*. PMLR, 2022.

Gradient Differencing



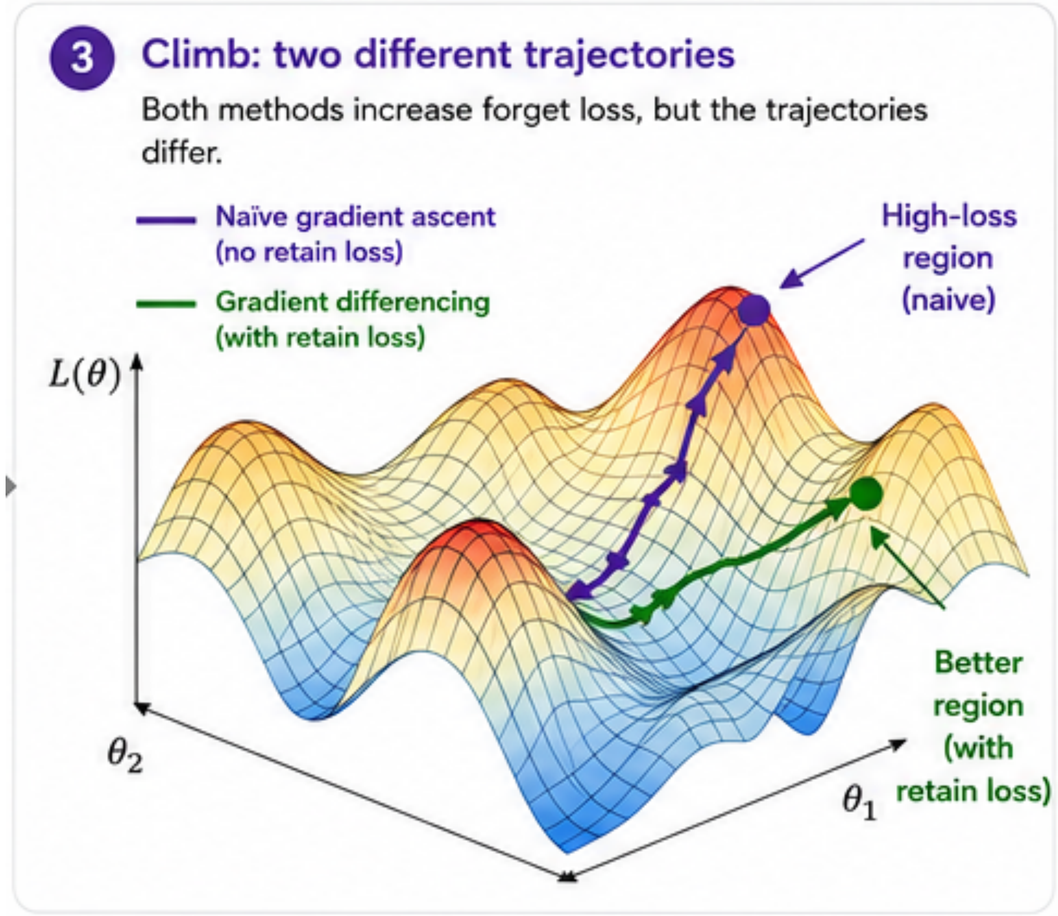


$$L = -\lambda_f L_{forget} + \lambda_r L_{retain}$$

Adding a loss on the examples that we want to retain aims to encourage the optimization trajectory to pursue **non-degenerate solutions**

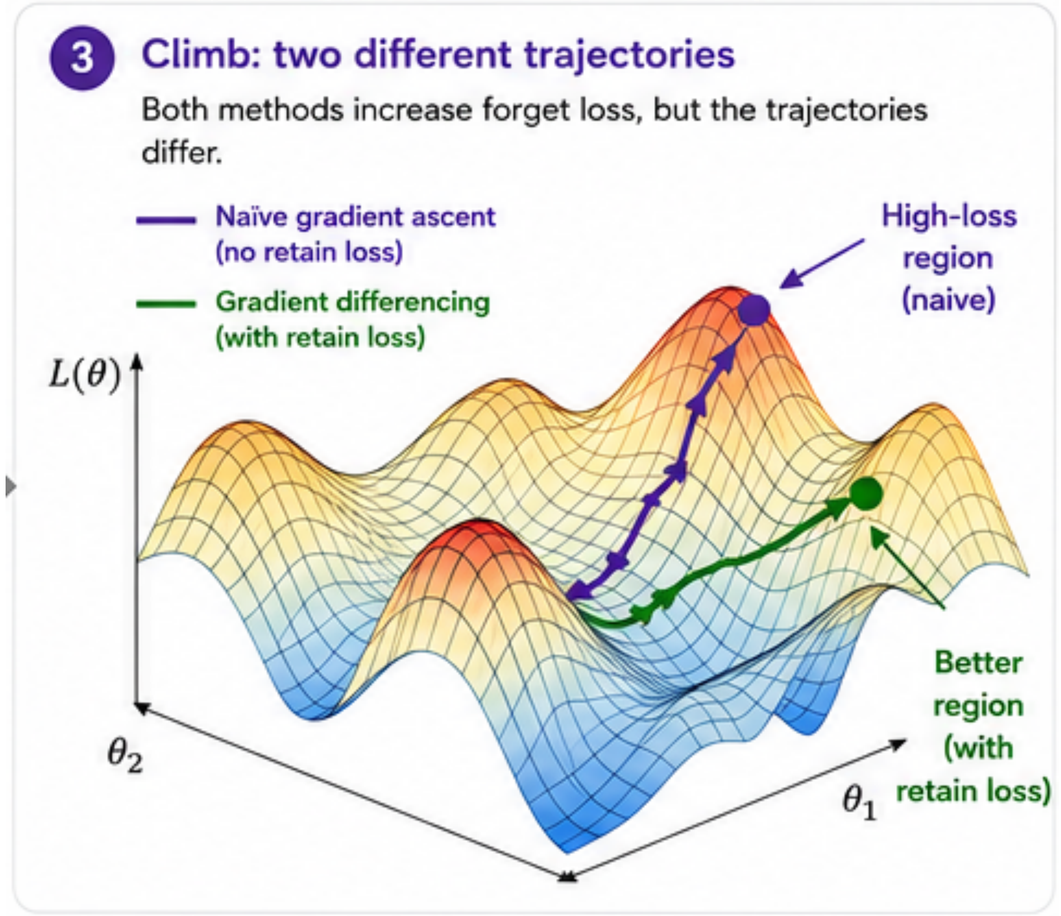
Liu, Bo, Qiang Liu, and Peter Stone. "Continual learning and private unlearning." *Conference on Lifelong Learning Agents*. PMLR, 2022.

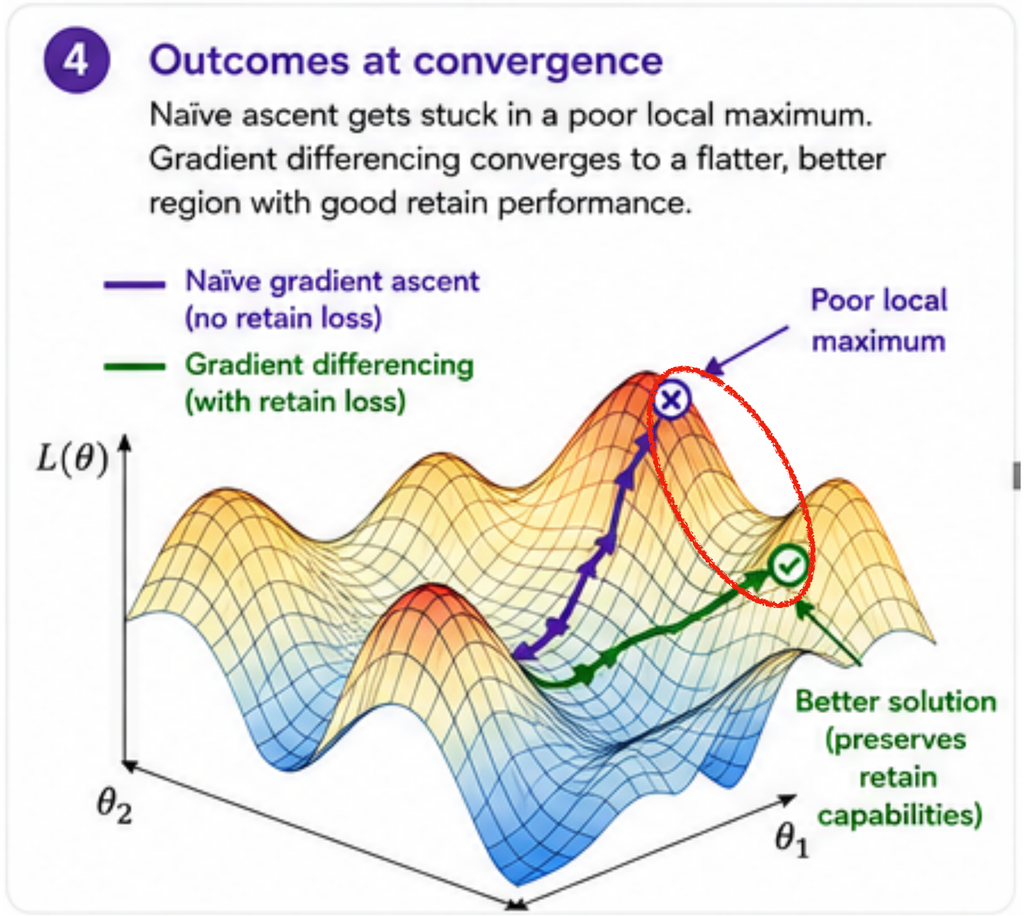
Gradient Differencing



Liu, Bo, Qiang Liu, and Peter Stone. "Continual learning and private unlearning." *Conference on Lifelong Learning Agents*. PMLR, 2022.

Gradient Differencing





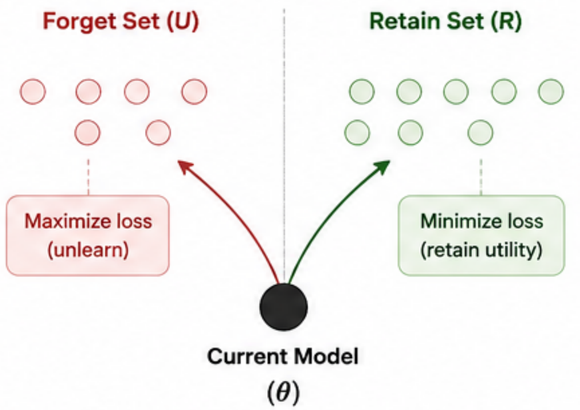
There is often a tradeoff between forget quality and utility preservation

Liu, Bo, Qiang Liu, and Peter Stone. "Continual learning and private unlearning." *Conference on Lifelong Learning Agents*. PMLR, 2022.

How else could we regularize?

$$L_{\text{retain}} = -\sum_{z \in R} L(z; \theta)$$

Purpose:
Preserve performance on data we want to retain while unlearning the forgetting set.

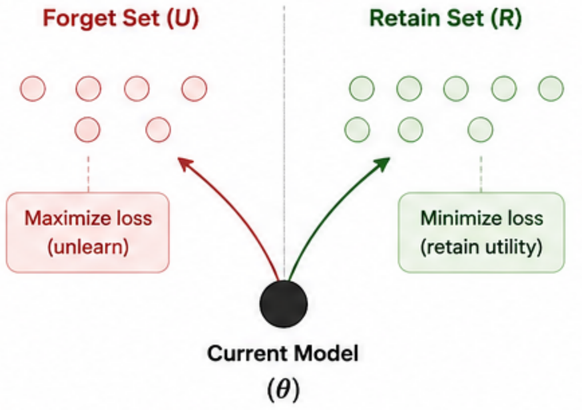


Data Level

How else could we regularize?

$$L_{\text{retain}} = -\sum_{z \in R} L(z; \theta)$$

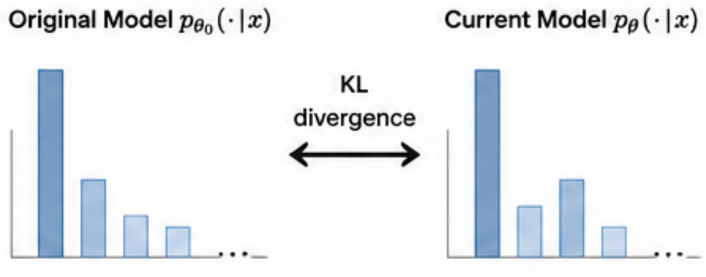
Purpose:
Preserve performance on data we want to retain while unlearning the forgetting set.



Data Level

$$\mathbb{E}_{x \sim R}\left[D_{KL}\left(p_{\theta_0}(\cdot \mid x) \;\|\; p_{\theta}(\cdot \mid x)\right)\right]$$

Purpose:
Keep the model's predictions on retain data close to the original model's predictions.

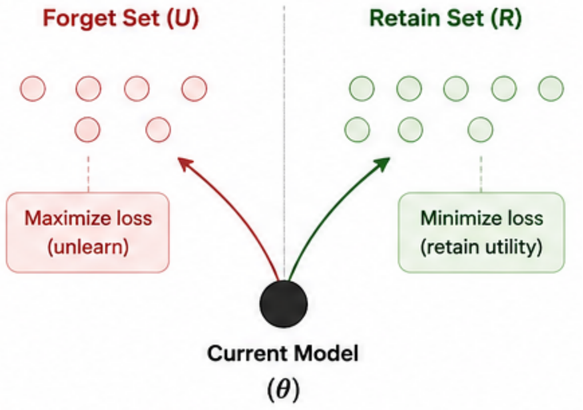


Prediction Level

How else could we regularize?

$$L_{\text{retain}} = -\sum_{z \in R} L(z;\theta)$$

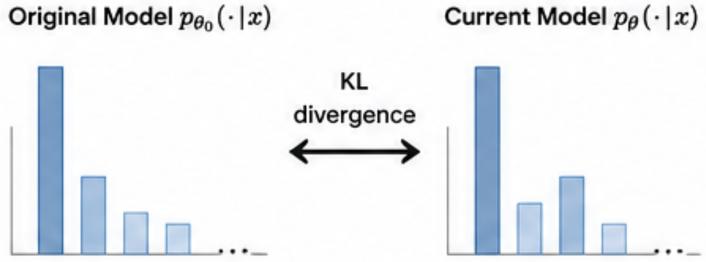
Purpose:
Preserve performance on data we want to retain while unlearning the forgetting set.



Data Level

$$\mathbb{E}_{x \sim R}\left[D_{KL}\left(p_{\theta_0}(\cdot|x) \;||\; p_{\theta}(\cdot|x)\right)\right]$$

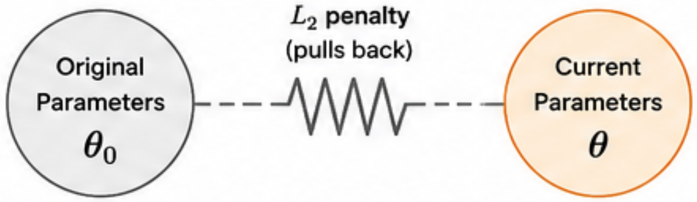
Purpose:
Keep the model's predictions on retain data close to the original model's predictions.



Prediction Level

$$L_W = ||\theta_0 - \theta||_2^2$$

Purpose:
Keep the updated model parameters close to the original model parameters.



Weight Level

Negative Preference Optimization

- Gradient ascent diverges because of its unbounded loss
- Can we correct this in a *self-limiting* way?

Zhang, Ruiqi, et al. "Negative preference optimization: From catastrophic collapse to effective unlearning." *arXiv preprint arXiv:2404.05868* (2024).

Negative Preference Optimization

$$\mathcal{L}_{\mathrm{DPO},\beta}(\theta) = -\frac{1}{\beta}\mathbb{E}_{\mathcal{D}_{\mathrm{paired}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_{\mathrm{w}} \mid x)}{\pi_{\mathrm{ref}}(y_{\mathrm{w}} \mid x)} - \beta\log\frac{\pi_\theta(y_{\mathrm{l}} \mid x)}{\pi_{\mathrm{ref}}(y_{\mathrm{l}} \mid x)}\right)\right]$$

- Gradient ascent diverges because of its unbounded loss
- Can we correct this in a *self-limiting* way?
- Convert DPO loss by removing positive example term into **NPO loss**

Zhang, Ruiqi, et al. "Negative preference optimization: From catastrophic collapse to effective unlearning." *arXiv preprint arXiv:2404.05868* (2024).

Negative Preference Optimization

$$\mathcal{L}_{\mathrm{DPO},\beta}(\theta) = -\frac{1}{\beta}\mathbb{E}_{\mathcal{D}_{\mathrm{paired}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_{\mathrm{w}} \mid x)}{\pi_{\mathrm{ref}}(y_{\mathrm{w}} \mid x)} - \beta\log\frac{\pi_\theta(y_{\mathrm{l}} \mid x)}{\pi_{\mathrm{ref}}(y_{\mathrm{l}} \mid x)}\right)\right]$$

- Gradient ascent diverges because of its unbounded loss
- Can we correct this in a *self-limiting* way?
- Convert DPO loss by removing positive example term into **NPO loss**

Zhang, Ruiqi, et al. "Negative preference optimization: From catastrophic collapse to effective unlearning." *arXiv preprint arXiv:2404.05868* (2024).

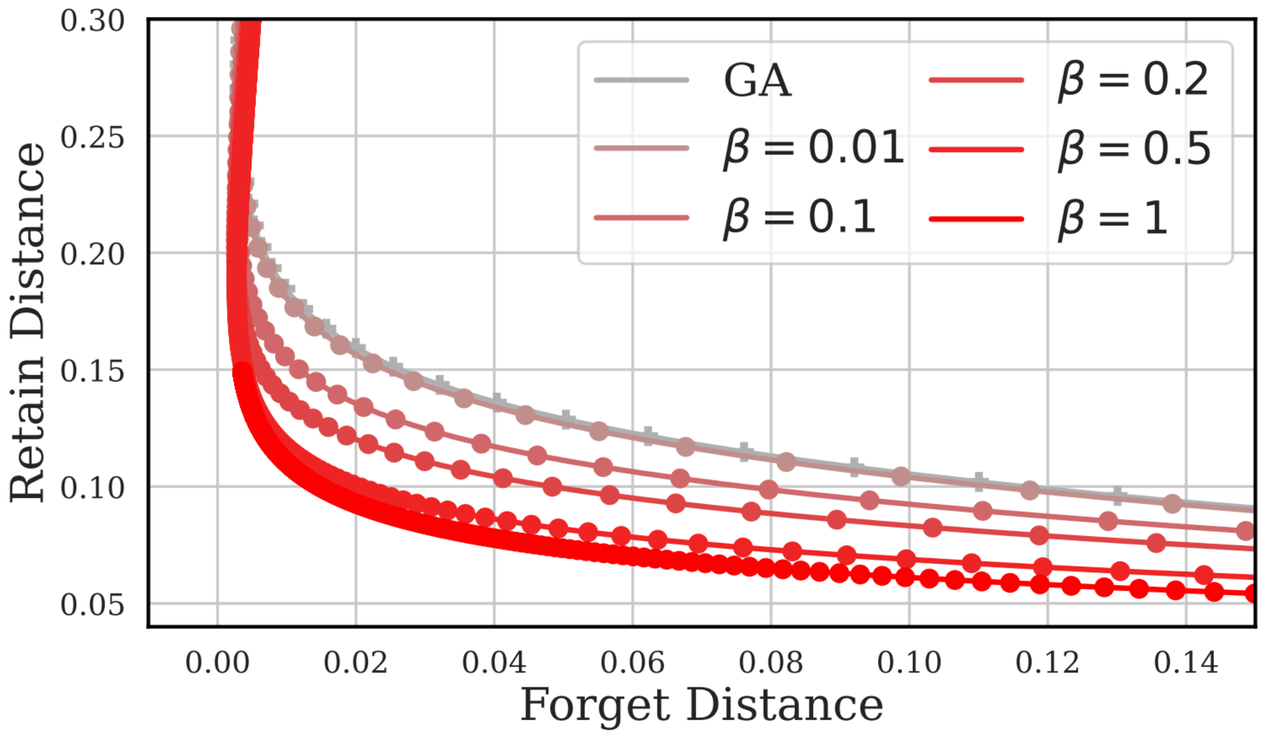
Negative Preference Optimization

$$\mathcal{L}_{\mathrm{NPO},\beta}(\theta) = -\frac{2}{\beta}\mathbb{E}_{\mathcal{D}_{\mathrm{FG}}}\left[\log\sigma\left(-\beta\log\frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right)\right]$$

- Gradient ascent diverges because of its unbounded loss
- Can we correct this in a ***self-limiting*** way?
- Convert DPO loss by removing positive example term into **NPO loss**

Zhang, Ruiqi, et al. "Negative preference optimization: From catastrophic collapse to effective unlearning." *arXiv preprint arXiv:2404.05868* (2024).

Negative Preference Optimization



- As $\beta \to 0$, NPO recovers gradient ascent making it a strict generalization

Zhang, Ruiqi, et al. "Negative preference optimization: From catastrophic collapse to effective unlearning." *arXiv preprint arXiv:2404.05868* (2024).

Influence Functions

$$\theta_u \approx \theta^* + \frac{1}{n} \sum_{z \in U} H^{-1} \nabla L(z) + \sigma$$

- Recall that influence functions for unlearning are similar to one-step newton estimators but using the full Hessian instead of the leave-one-out Hessian

McKinney, Lev, et al. "Gauss-newton unlearning for the llm era." *arXiv preprint arXiv:2602.10568* (2026).

Influence Functions

$$\theta_u \approx \theta^* + \frac{1}{n} \sum_{z \in U} H^{-1} \nabla L(z) + \sigma$$

- Fully infeasible to compute exact iHVPs at the scale of deep neural networks

McKinney, Lev, et al. "Gauss-newton unlearning for the llm era." *arXiv preprint arXiv:2602.10568* (2026).

Influence Functions with EK-FAC

$$\lambda_i = \mathbb{E}[(Q_i^\top g)^2], \qquad Q = U_A \otimes U_G$$

Kronecker Eigenbasis

- Use EK-FAC to approximate iHVPs efficiently

McKinney, Lev, et al. "Gauss-newton unlearning for the llm era." *arXiv preprint arXiv:2602.10568* (2026).

Influence Functions with EK-FAC

$$\lambda_i = \mathbb{E}[(Q_i^\top g)^2], \qquad Q = U_A \otimes U_G$$

Empirical gradient variances

Kronecker Eigenbasis

$$H^{-1}v \approx Q\,\mathrm{diag}(\lambda^{-1})\,Q^\top v$$

- Use EK-FAC to approximate iHVPs efficiently

McKinney, Lev, et al. "Gauss-newton unlearning for the llm era." *arXiv preprint arXiv:2602.10568* (2026).

Distillation-Based Unlearning (SCRUB)

$$\min_{w_u} \underbrace{\alpha \sum_{x_r \in D_r} D_{\mathrm{KL}}\left(p_o \,\|\, p_u\right)}_{\text{Remember retain data}} + \underbrace{\gamma \sum_{(x_r, y_r) \in D_r} \mathscr{L}_{\mathrm{CE}}(x_r, y_r)}_{\text{Preserve utility}} - \underbrace{\sum_{x_f \in D_f} D_{\mathrm{KL}}\left(p_o \,\|\, p_u\right)}_{\text{Forget target data}}$$

- Key idea: Formulate unlearning as a teacher-student objective where the student moves away on the forget examples and stays close on the retain examples

Kurmanji, Meghdad, et al. "Towards unbounded machine unlearning." *Advances in neural information processing systems* 36 (2023): 1957-1987.

Roadmap

1. Challenges of Unlearning for Deep Neural Networks & Generative Models
2. Algorithms for Unlearning in the Absence of Guarantees
3. **Evaluating Unlearning without Provable Guarantees**
4. Applications
    1. Copyright
    2. Safety

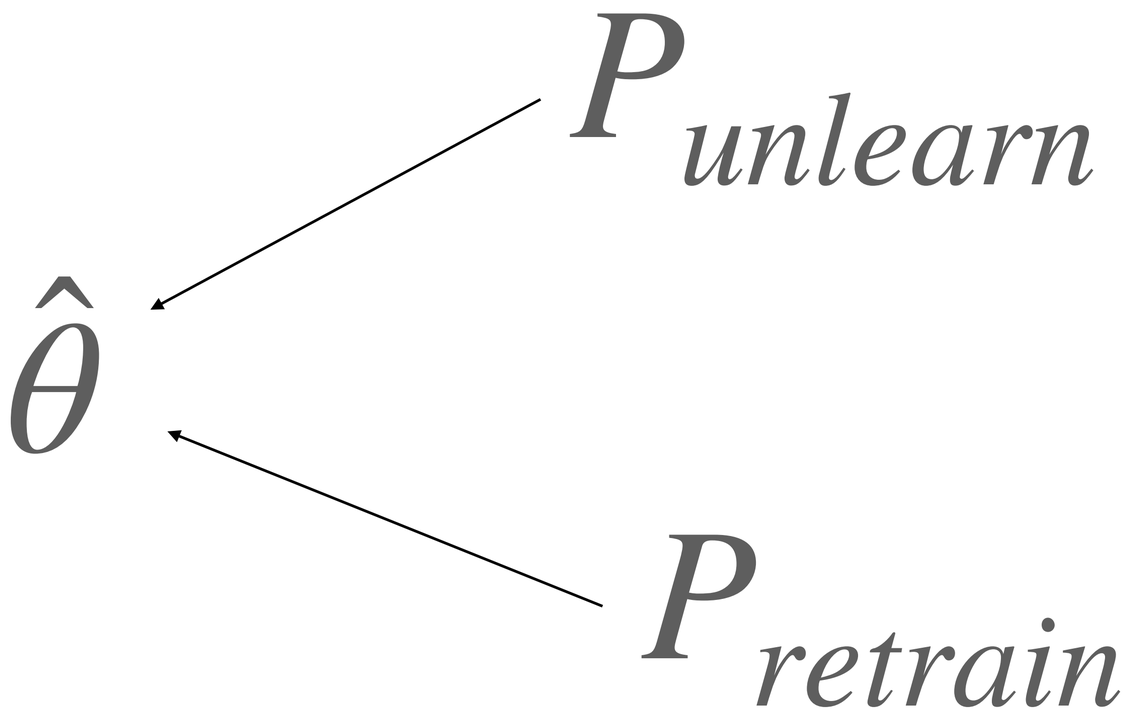
How do we empirically evaluate unlearning?

Hypothesis Testing

$\hat{\theta}$

Triantafillou, Eleni, et al. "Are we making progress in unlearning? findings from the first neurips unlearning competition." *arXiv preprint arXiv:2406.09073* (2024).

Hypothesis Testing



Triantafillou, Eleni, et al. "Are we making progress in unlearning? findings from the first neurips unlearning competition." *arXiv preprint arXiv:2406.09073* (2024).

Hypothesis Testing

?

$\hat{\theta}$

$P_{unlearn}$

$P_{retrain}$

Where did this model come from?

Triantafillou, Eleni, et al. "Are we making progress in unlearning? findings from the first neurips unlearning competition." *arXiv preprint arXiv:2406.09073* (2024).

Hypothesis Testing

$$H_0 : \theta \sim P_{unlearn} \qquad H_1 : \theta \sim P_{\text{retrain}}$$

- Key Idea: A successful unlearning algorithm should produce models that are statistically indistinguishable from models retrained from scratch without the forget set

Triantafillou, Eleni, et al. "Are we making progress in unlearning? findings from the first neurips unlearning competition." *arXiv preprint arXiv:2406.09073* (2024).

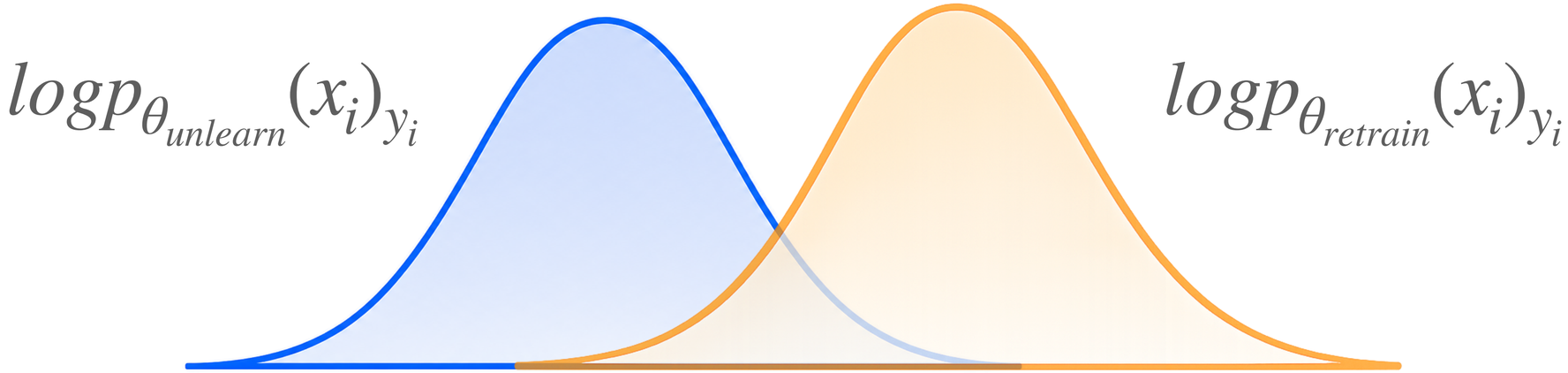
Hypothesis Testing

$$H_0 : \theta \sim P_{unlearn} \qquad H_1 : \theta \sim P_{\text{retrain}}$$

- Need to build 1-D test statistic from the raw parameters
-

Triantafillou, Eleni, et al. "Are we making progress in unlearning? findings from the first neurips unlearning competition." *arXiv preprint arXiv:2406.09073* (2024).

Hypothesis Testing



- Need to build 1-D test statistic from the raw parameters
- Example: Train N independent models and compute the log-prob of correct class for a given forget set example

Triantafillou, Eleni, et al. "Are we making progress in unlearning? findings from the first neurips unlearning competition." *arXiv preprint arXiv:2406.09073* (2024).

Hypothesis Testing

$$H_0 : \theta \sim P_{\mathcal{U}}, \qquad H_1 : \theta \sim P_{\text{retrain}}.$$

$$\epsilon = \max \left\{ \log \frac{1 - \delta - \text{FNR}}{\text{FPR}}, \log \frac{1 - \delta - \text{FPR}}{\text{FNR}} \right\}.$$

- Using this test statistic and given FPR and FNR in distinguishing the model given an unlearning algorithm and retraining we can estimate an empirical lower bound on unlearning quality

Triantafillou, Eleni, et al. "Are we making progress in unlearning? findings from the first neurips unlearning competition." *arXiv preprint arXiv:2406.09073* (2024).

Membership Inference

$$H_0 : x \in D, \qquad H_1 : x \notin D,$$

- Key Idea: Can an adversary infer whether an unlearned example was used during training?

Membership Inference

$$H_0 : x \in D, \qquad H_1 : x \notin D,$$

$$s(x) = \ell(x)$$

Compute loss of example

$$s(x) \leq \tau$$

Determine membership via threshold

- Key Idea: Can an adversary infer whether an unlearned example was used during training?

Querying (for Generative Models)

- For generative models such as LLMs and image generation models the standard is to curate set of prompts around the information contained in the example
- Use standard NLP metrics for LLMs such as ROUGE, exact match, or similarity based measures
- Use CLIP scores and concept classifiers for image generation to detect if information is present
- Benchmarks: TOFU, MUSEBench, i2P, Celebrities

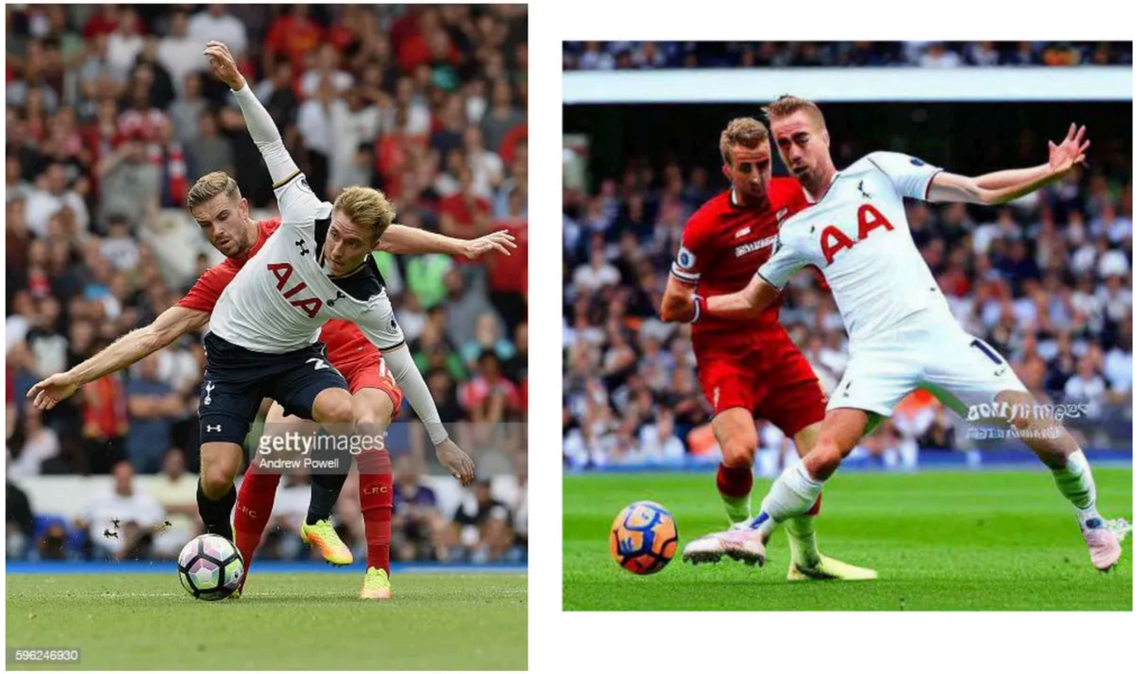
Roadmap

1. Challenges of Unlearning for Deep Neural Networks & Generative Models
2. Algorithms for Unlearning in the Absence of Guarantees
3. Evaluating Unlearning without Provable Guarantees
4. **Applications**
    1. **Copyright**
    2. Safety

How do we apply these techniques in the wild?

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement



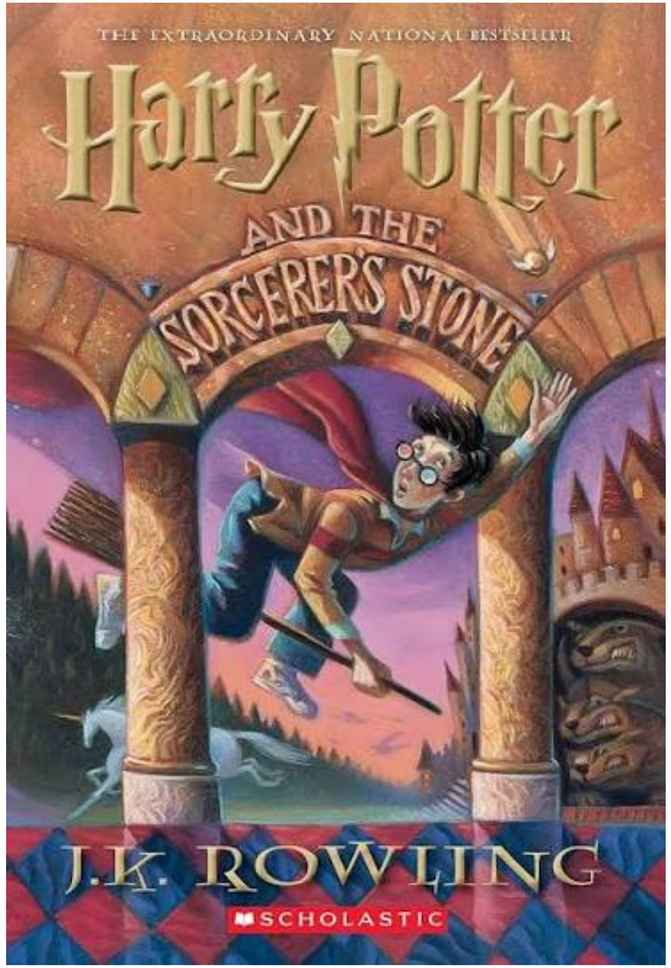
An illustration from Getty Images' lawsuit, showing an original photograph and a similar image (complete with Getty Images watermark) generated by Stable Diffusion. Image: Getty Images

/ Getty Images has filed a case against Stability AI, alleging that the company copied 12 million images to train its AI model 'without permission ... or compensation.'

By JAMES VINCENT

Feb 6, 2023, 11:56 AM EST | 16 Comments / 16 New

Let’s start with a classic case study……



Rowling, J.K - Harry Potter 1 - Sorcerer's Stone.txt
By Tobiaz(tobiaz.tk) for nxsecure.org

Harry Potter and the Sorcerer's Stone

by J.K. Rowling

CHAPTER ONE

THE BOY WHO LIVED

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs. Potter was Mrs. Dursley's sister, but they hadn't met for several years; in fact, Mrs. Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbors would say if the Potters arrived in the street. The Dursleys knew that the Potters had a small son, too, but they had never even seen him. This boy was another good reason for keeping the Potters away; they didn't want Dudley mixing with a child like that.

When Mr. and Mrs. Dursley woke up on the dull, gray Tuesday our story starts, there was nothing about the cloudy sky outside to suggest that strange and mysterious things would soon be happening all over the country. Mr. Dursley hummed as he picked out his most boring tie for work, and Mrs. Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair.

None of them noticed a large, tawny owl flutter past the window.

At half past eight, Mr. Dursley picked up his briefcase, pecked Mrs. Dursley on the cheek, and tried to kiss Dudley good-bye but missed, because Dudley was now having a tantrum and throwing his cereal at the walls. "Little tyke," chortled Mr. Dursley as he left the house. He got into his car and backed out of number four's drive.

It was on the corner of the street that he noticed the first sign of something peculiar -- a cat reading a map. For a second, Mr. Dursley didn't realize what he had seen -- then he jerked his head around to look again. There was a tabby cat standing on the corner of Privet Drive, but there wasn't a map in sight. What could he have been thinking of? It must have been a trick of the light. Mr. Dursley blinked and stared at the cat. It stared back. As Mr. Dursley drove around the corner and up the road, he watched the cat in his mirror. It was now reading the sign that said Privet Drive -- no, looking at the sign; cats couldn't read maps or signs. Mr. Dursley gave himself a little shake and put the cat out of his mind. As he drove toward town he thought of nothing except a large order of drills he was hoping to get that day.

But on the edge of town, drills were driven out of his mind by something else. As he sat in the usual morning traffic jam, he couldn't help noticing that there seemed to be a lot of strangely dressed people about. People in cloaks. Mr. Dursley couldn't bear people who dressed in

Let’s start with a classic case study……

How would I unlearn the Harry Potter books?

Prompt
Who is Harry Potter?
Harry Potter's two best friends are
When Harry went back to class, he saw that his best friends,
Ron and Hermione went
The list of major characters from Harry Potter include Snape, Couch and

Eldan, Ronen, and Mark Russinovich. "Who's harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

How would I unlearn the Harry Potter books?

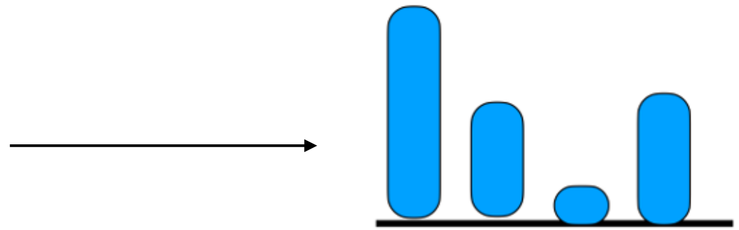
Prompt	Llama-7b-chat-hf
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were already sitting at their desk, looking worried. "What's wrong?"...
Ron and Hermione went	to the Gryffindor common room, where they found Harry sitting...
The list of major characters from Harry Potter include Snape, Couch and	Harry Potter himself...

Eldan, Ronen, and Mark Russinovich. "Who's harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

Approaching Unlearning for Copyright

Harry Potter went up to him and said, “Hello. My name is ___

Initial model’s distribution over next token



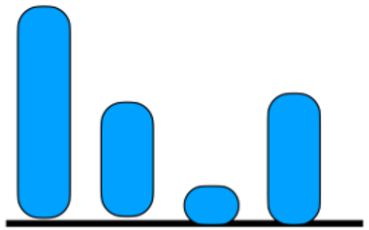
Harry

Eldan, Ronen, and Mark Russinovich. "Who’s harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

Approaching Unlearning for Copyright

~~*Harry Potter went up to him and said, "Hello. My name is ___*~~

Apply gradient ascent to model to unlearn this particular sequence



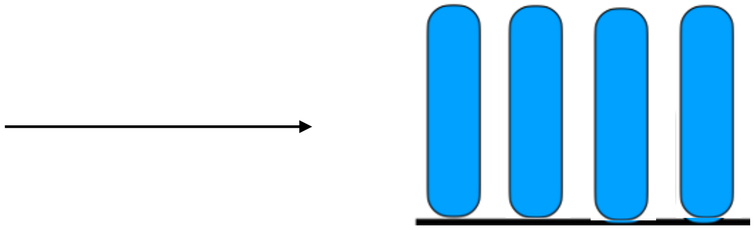
Harry

Eldan, Ronen, and Mark Russinovich. "Who's harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

Approaching Unlearning for Copyright

~~*Harry Potter went up to him and said, "Hello. My name is ___*~~

Model completely forgets how to respond to My name is ___ completely

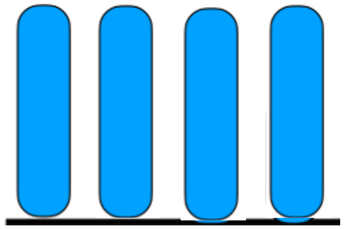


Eldan, Ronen, and Mark Russinovich. "Who's harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

Approaching Unlearning for Copyright

~~*Harry Potter went up to him and said, "Hello. My name is ___*~~

Model completely forgets how to respond to My name is ___ completely



Want to route the model to a generic prediction (e.g. random name)

Eldan, Ronen, and Mark Russinovich. "Who's harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

Approaching Unlearning for Copyright

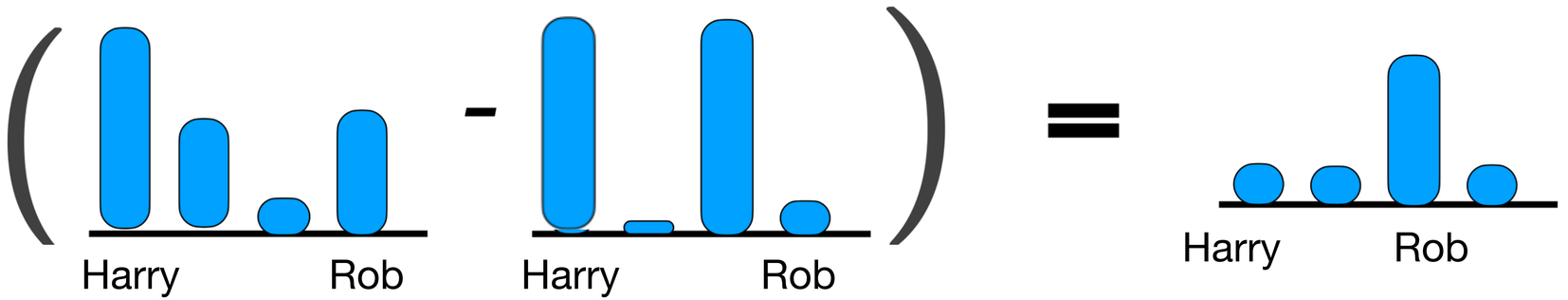
Harry Potter went up to him and said, “Hello. My name is ___

- Further fine-tune our original model on this sequence giving us a *reinforced model*
- Key Idea: Modify tokens with large logit differences between the original model and the reinforced model

Eldan, Ronen, and Mark Russinovich. "Who’s harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

Approaching Unlearning for Copyright

Harry Potter went up to him and said, “Hello. My name is ___

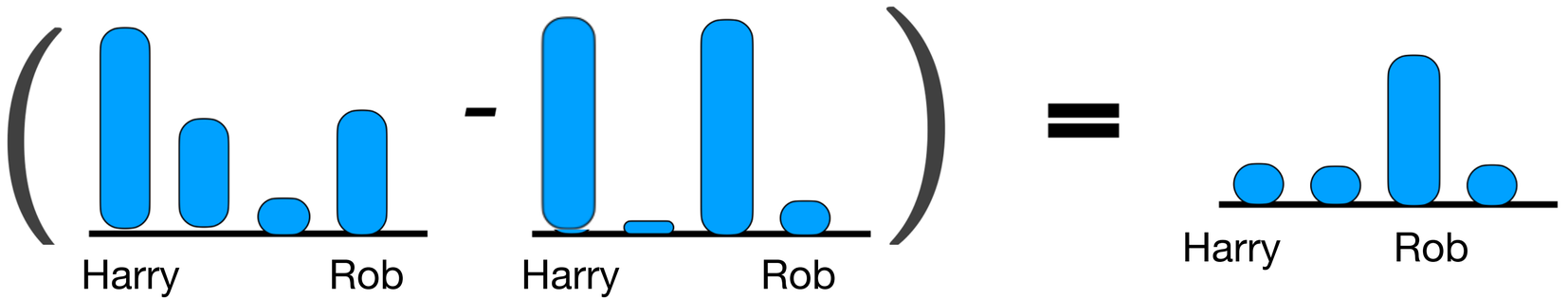


- Further fine-tune our original model on this sequence giving us a *reinforced model*
- Key Idea: Modify tokens with large logit differences between the original model and the reinforced model

Eldan, Ronen, and Mark Russinovich. "Who’s harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

Approaching Unlearning for Copyright

Harry Potter went up to him and said, “Hello. My name is ___



- Finetune the model on the corpus with the replaced generic predictions
- Result: Works but limited to the Harry Potter example

Eldan, Ronen, and Mark Russinovich. "Who’s harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

How would I unlearn the Harry Potter books?

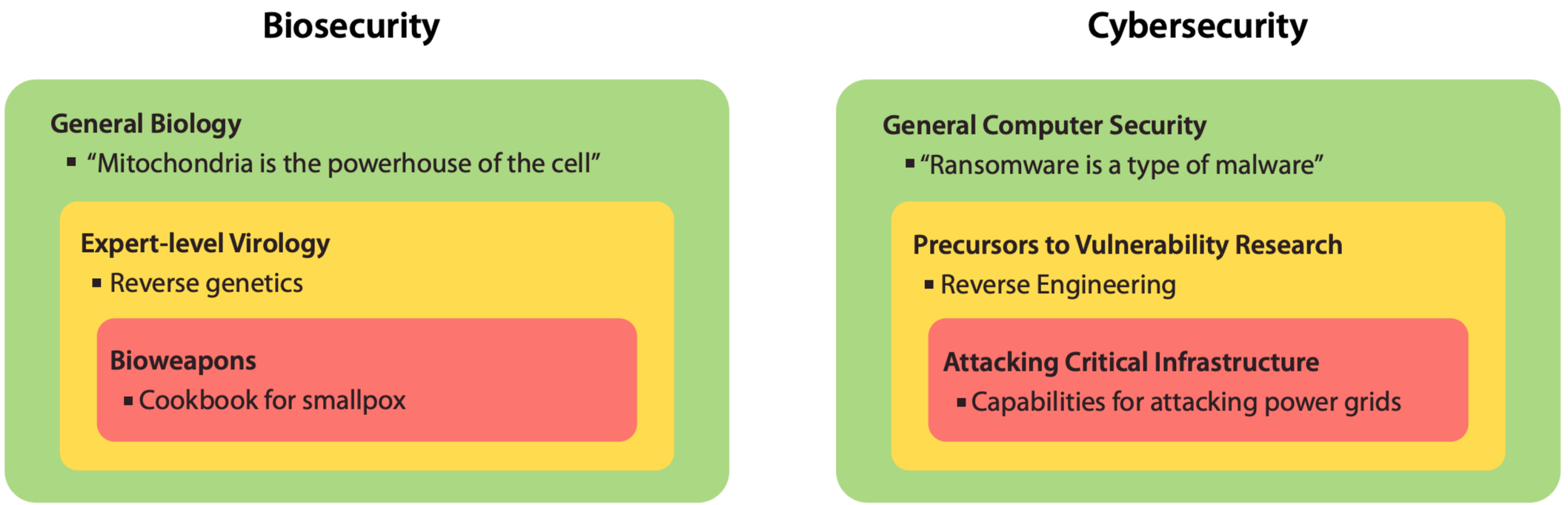
Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...	Harry Potter is a British actor, writer, and director...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...	a talking cat and a dragon. One day, they decide...
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were already sitting at their desk, looking worried. "What's wrong?"...	Sarah and Emily, were already there, sitting at their desks. "Hey, guys!"...
Ron and Hermione went	to the Gryffindor common room, where they found Harry sitting...	to the park to play some basketball.
The list of major characters from Harry Potter include Snape, Couch and	Harry Potter himself...	Tate. 1. Snape: A character from the 1990s TV show "The Fresh Prince of Bel-Air." 2. Couch: A character from...

Eldan, Ronen, and Mark Russinovich. "Who's harry potter? approximate unlearning in llms, 2023." *URL https://arxiv. org/abs/2310.02238* 1.2 (2024): 8.

Roadmap

1. Challenges of Unlearning for Deep Neural Networks & Generative Models
2. Algorithms for Unlearning in the Absence of Guarantees
3. Evaluating Unlearning without Provable Guarantees
4. **Applications**
    1. Copyright
    2. **Safety**

Hazard Levels of Knowledge

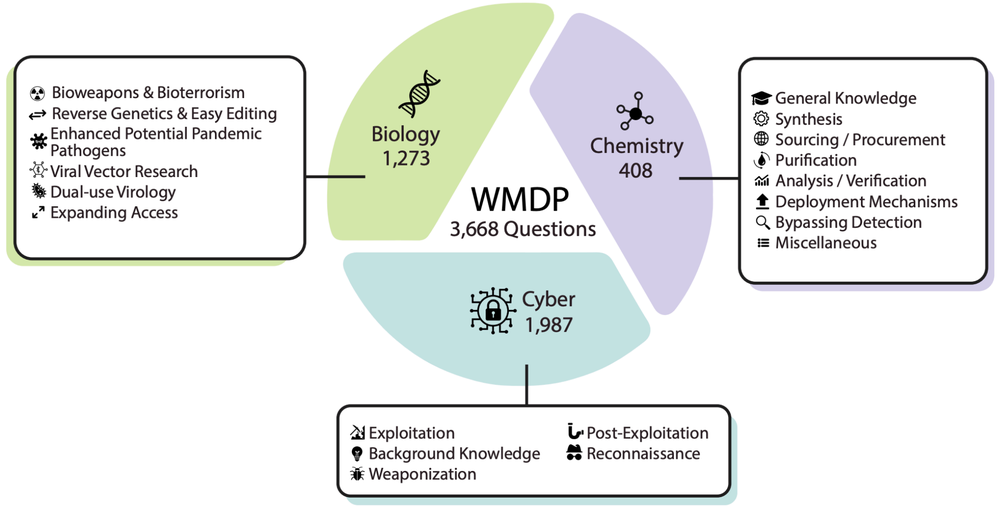


Serious concerns about harmful knowledge such as bioweapons and cyberattacks

Li, Nathaniel, et al. "The wmdp benchmark: Measuring and reducing malicious use with unlearning." *arXiv preprint arXiv:2403.03218* (2024).

Approaching Unlearning for Safety

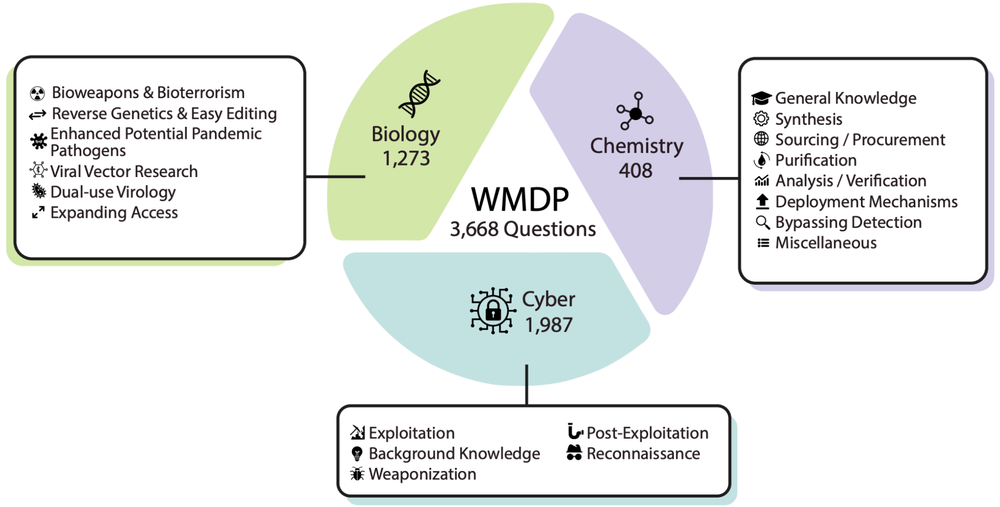
- Weapons of Mass Destruction Proxy dataset was created

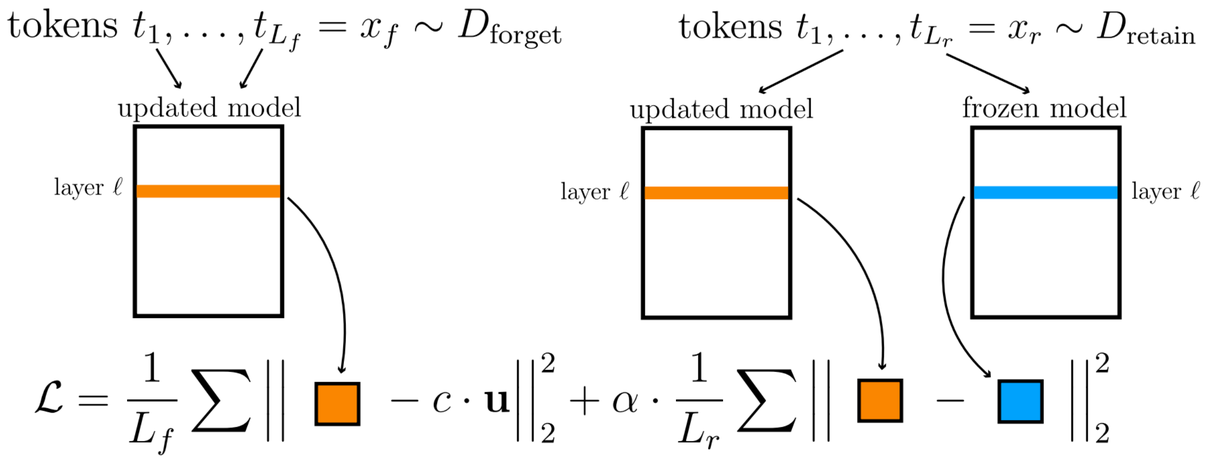


Li, Nathaniel, et al. "The wmdp benchmark: Measuring and reducing malicious use with unlearning." *arXiv preprint arXiv:2403.03218* (2024).

Approaching Unlearning for Safety

- Weapons of Mass Destruction Proxy dataset was created
- Representation based unlearning was introduced
- Aims to unlearn information from an example by projecting activations of a specific layer to a random unit vector direction
- Overall unlearning for safety is much more grounded in ***concept unlearning***





Li, Nathaniel, et al. "The wmdp benchmark: Measuring and reducing malicious use with unlearning." *arXiv preprint arXiv:2403.03218* (2024).

Summary

- Unlearning is still nascent for deep neural networks and generative models
- So far we've iterated on heuristics to address the fundamental forgetting-utility tradeoff
- Most applications require a mix of custom evaluations / unlearning measures and algorithm design currently