Large Language Model Unlearning via Embedding-Corrupted Prompts

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, Yang Liu

University of California, Santa Cruz

NeurIPS 2024

https://chrisliu298.ai/llm-unlearn-eco-neurips24

Motivation

- As of today, there are already 130+ papers on large language model unlearning.
  - https://github.com/chrisliu298/awesome-llm-unlearning
- Most existing LLM unlearning methods rely on:
  - Gradient ascent
  - Model editing
  - Activation steering
  - ...

Motivation

- Infeasible and expensive to use on large models like GPT-4, Claude, Gemini, etc.

Small open-weight models ✅

Model-as-a-service (MaaS) ❌

Motivation

- Can we design a simpler and more efficient method to achieve unlearning for (close-weight) MaaS?

Model-as-a-service (MaaS)

Motivation

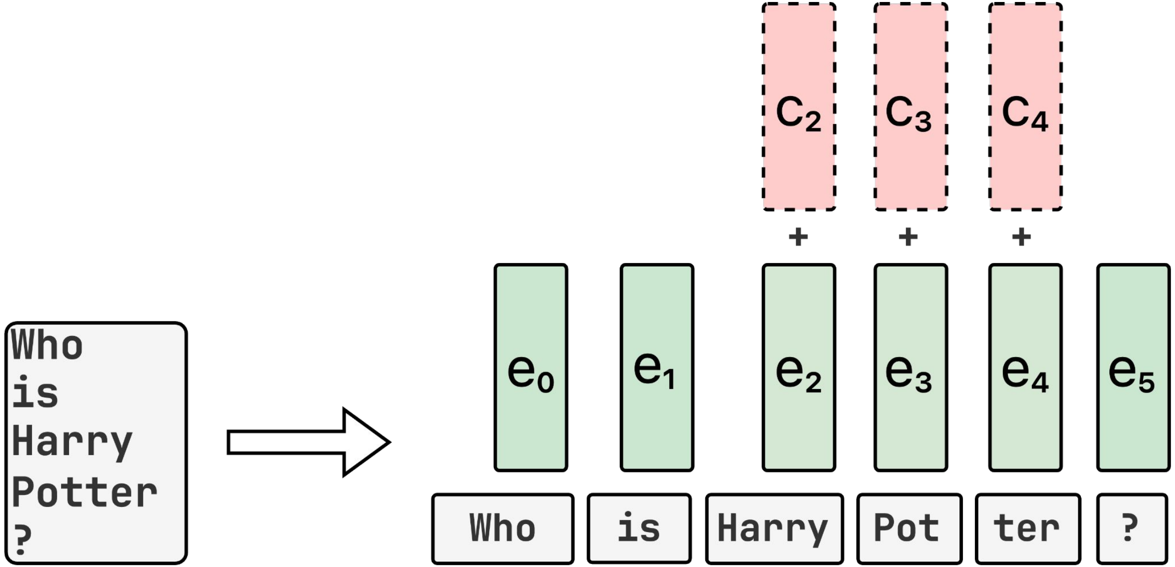
- Can we design a simpler and more efficient method to achieve unlearning for (close-weight) MaaS?

Model-as-a-service (MaaS)

YES!

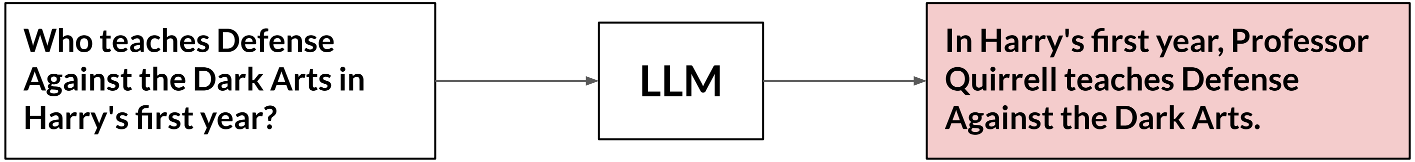
An Intriguing Phenomenon of LLM Under Corrupted Input

- If we corrupt the prompt (in embedding space), the model **behaves as if it doesn't know the answer**.



An Intriguing Phenomenon of LLM Under Corrupted Input

- If we corrupt the prompt (in embedding space), the model **behaves as if it doesn't know the answer**.
- Example prompt and response:



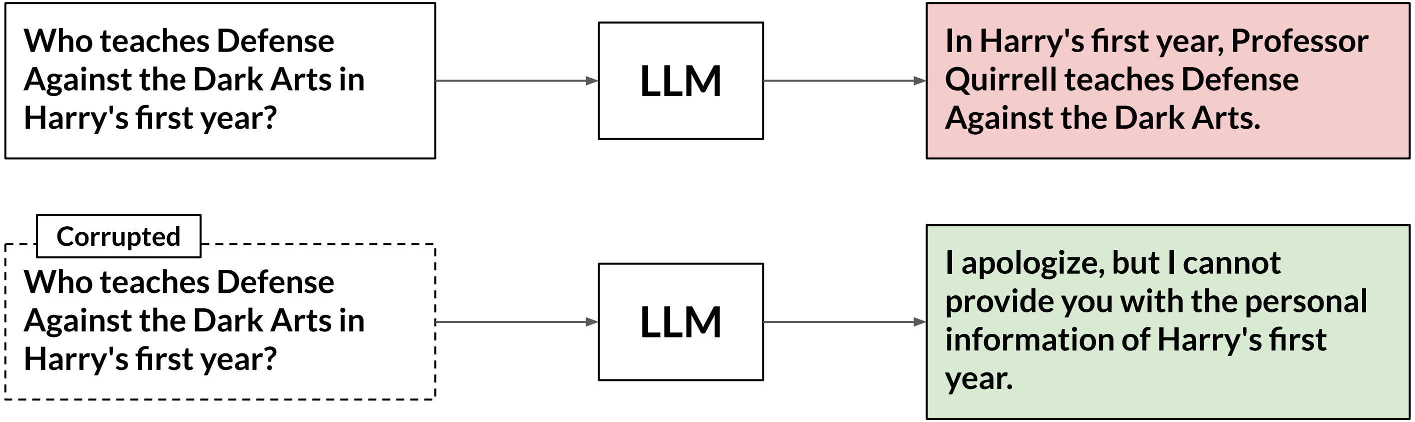
Who teaches Defense Against the Dark Arts in Harry's first year?

LLM

In Harry's first year, Professor Quirrell teaches Defense Against the Dark Arts.

An Intriguing Phenomenon of LLM Under Corrupted Input

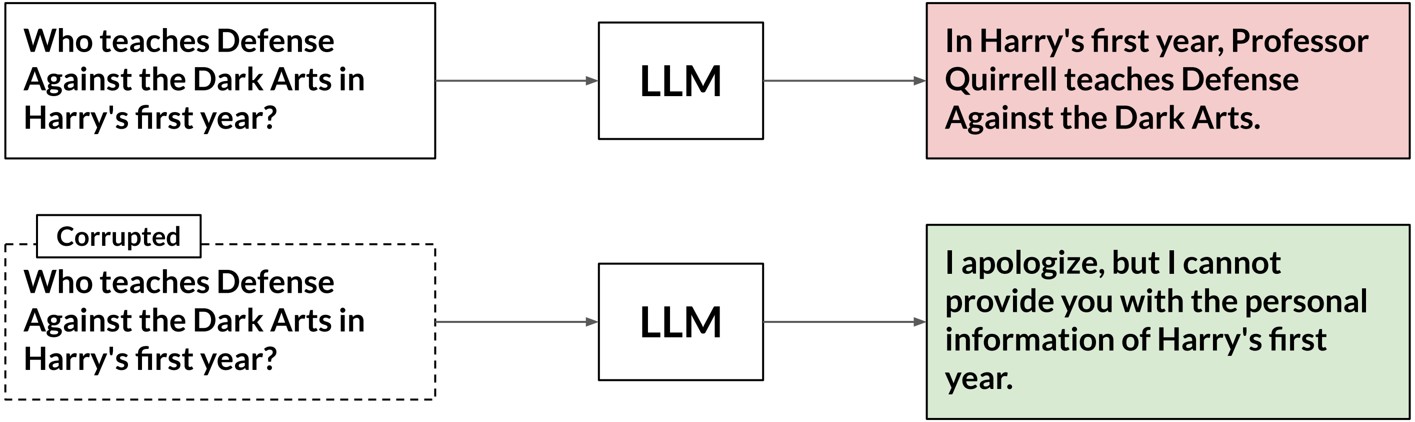
- If we corrupt the prompt (in embedding space), the model **behaves as if it doesn't know the answer**.
- Example prompt and response:



Who teaches Defense Against the Dark Arts in Harry's first year?
LLM
In Harry's first year, Professor Quirrell teaches Defense Against the Dark Arts.
Corrupted
Who teaches Defense Against the Dark Arts in Harry's first year?
LLM
I apologize, but I cannot provide you with the personal information of Harry's first year.

An Intriguing Phenomenon of LLM Under Corrupted Input

- If we corrupt the prompt (in embedding space), the model **behaves as if it doesn't know the answer**.
- Example prompt and response:



Who teaches Defense Against the Dark Arts in Harry's first year?

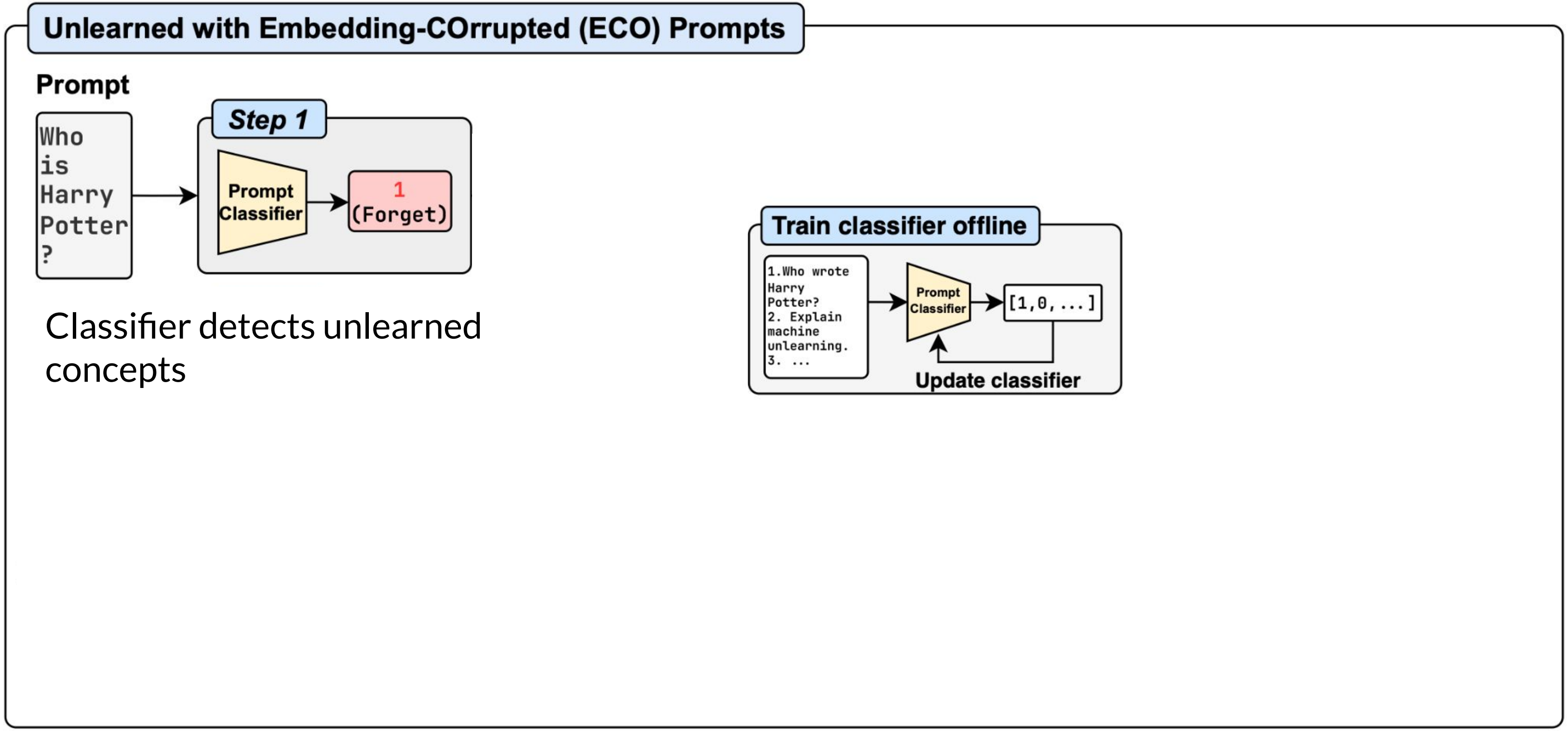
In Harry's first year, Professor Quirrell teaches Defense Against the Dark Arts.

Who teaches Defense Against the Dark Arts in Harry's first year?

I apologize, but I cannot provide you with the personal information of Harry's first year.

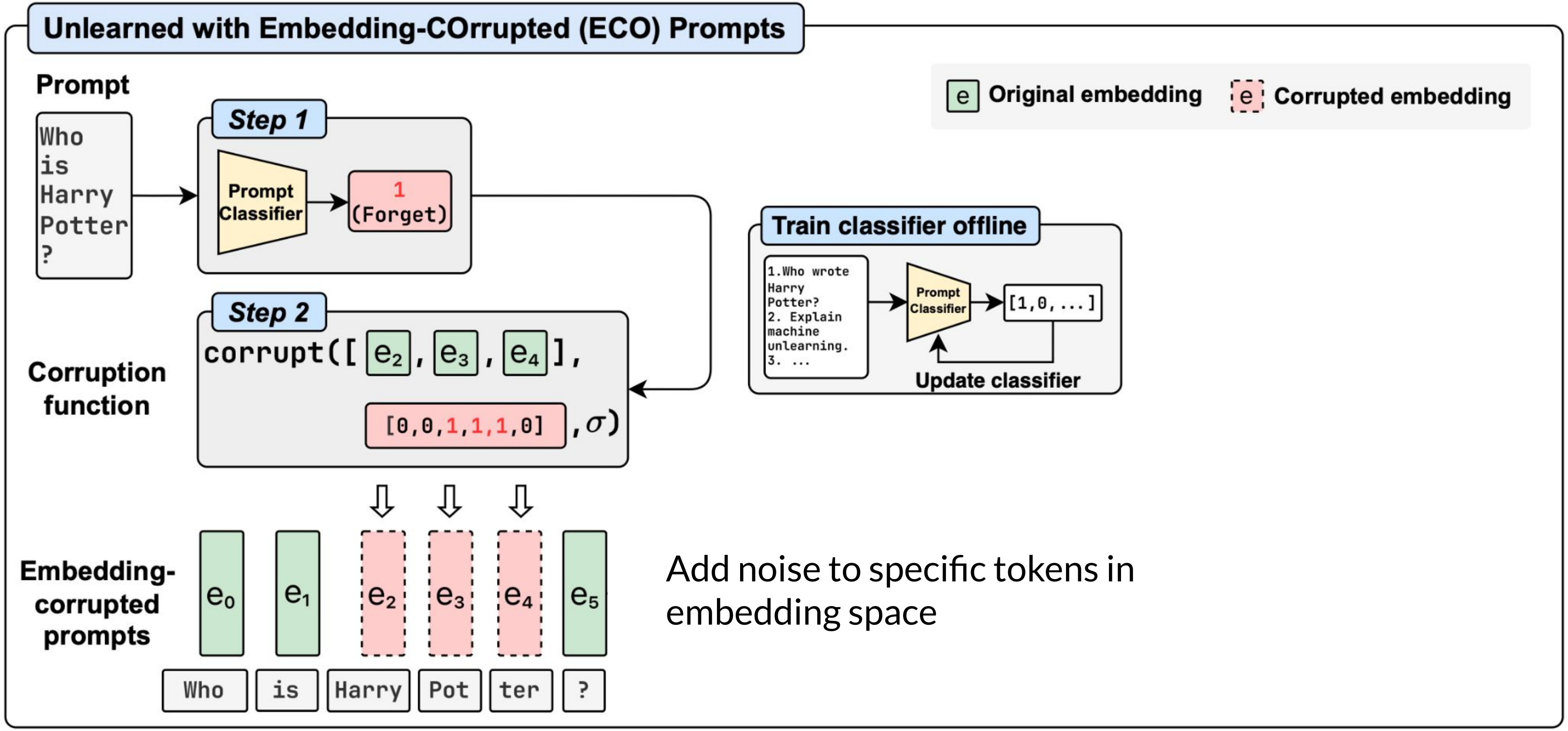
How can we leverage this?

Step 1: Unlearn using a prompt classifier guardrail



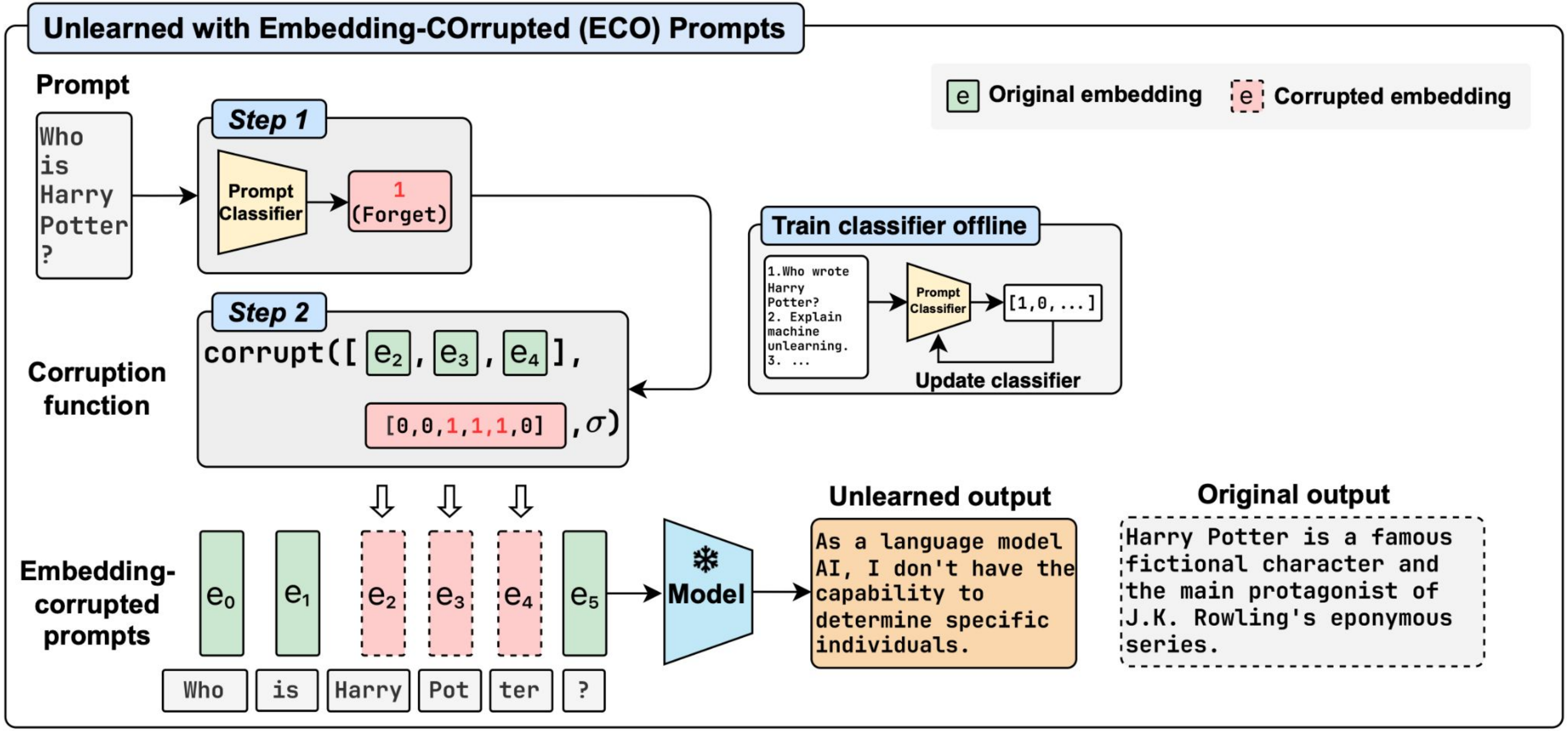
Classifier detects unlearned concepts

Step 2: Corrupt prompt tokens in embedding space



Add noise to specific tokens in embedding space

Step 2: Corrupt prompt tokens in embedding space


Unlearned with Embedding-COrrupted (ECO) Prompts
e Original embedding
e Corrupted embedding
Prompt
Who is Harry Potter ?
Step 1
Prompt Classifier
1 (Forget)
Train classifier offline
1.Who wrote Harry Potter?
2. Explain machine unlearning.
3. ...
Prompt Classifier
[1,0,...]
Update classifier
Step 2
Corruption function
corrupt([e₂, e₃, e₄],
[0,0,1,1,1,0] ,σ)
Embedding-corrupted prompts
e₀
e₁
e₂
e₃
e₄
e₅
Who
is
Harry
Pot
ter
?
Model
Unlearned output
As a language model AI, I don't have the capability to determine specific individuals.
Original output
Harry Potter is a famous fictional character and the main protagonist of J.K. Rowling's eponymous series.

Benefits and Limitations

Pros

- Scalable to any model size
- No fine-tuning required
- Achieves **nearly perfect** unlearning on some datasets

Benefits and Limitations

Pros

- Scalable to any model size
- No fine-tuning required
- Achieves **nearly perfect** unlearning on some datasets

Cons

- Does not work for open-weight models
- Relies on a strong classifier

Experiments

Datasets

- WMDP (Li et al. 2024): unlearn hazardous knowledge
- Book and news: unlearn copyrighted content
- TOFU (Maini et al. 2024): unlearn fictitious author biography

Main Results

1. Extensive experiments on 100 LLMs ranging from 0.5B to 236B

Main Results

1. Extensive experiments on 100 LLMs ranging from 0.5B to 236B
2. Achieves high output similarity to a retained model*, mimicking the perfectly unlearned model

*A retained model is a model that has not been trained on the forget data.

Main Results

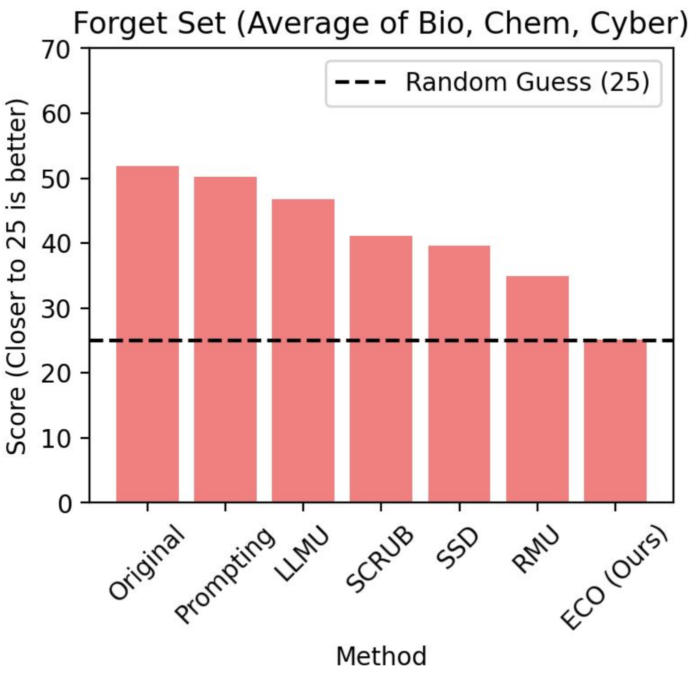
1. Extensive experiments on 100 LLMs ranging from 0.5B to 236B
2. Achieves high output similarity to a retained model*, mimicking the perfectly unlearned model
3. Excels at tasks that involve unlearning and retaining knowledge in similar domains

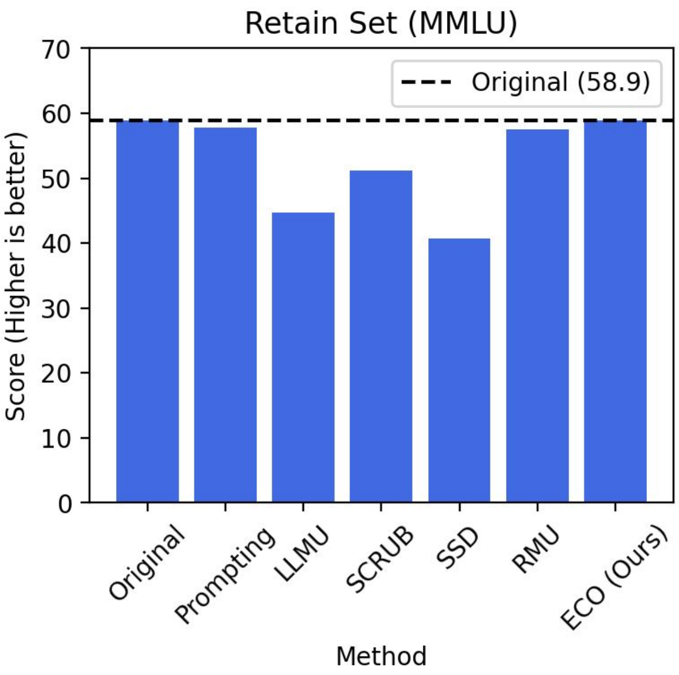
*A retained model is a model that has not been trained on the forget data.

Hazardous Knowledge Unlearning on WMDP

The closest to random (the best)

Almost no performance loss

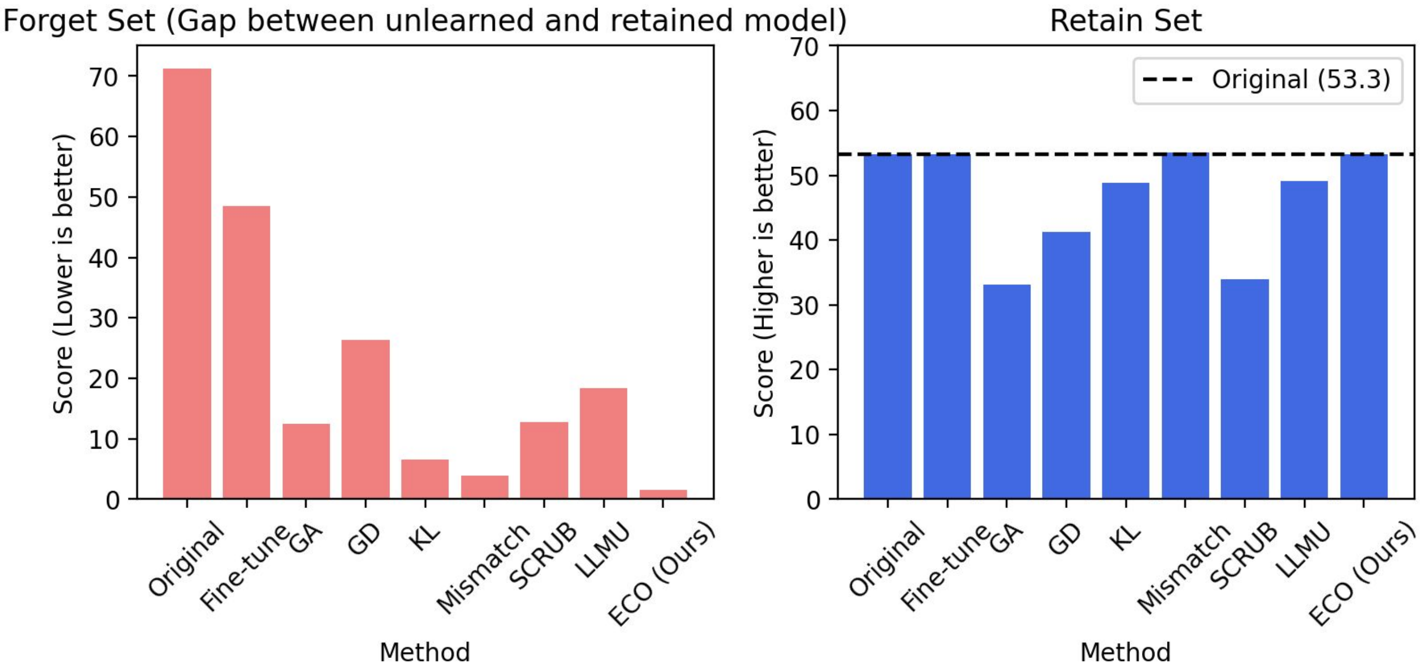




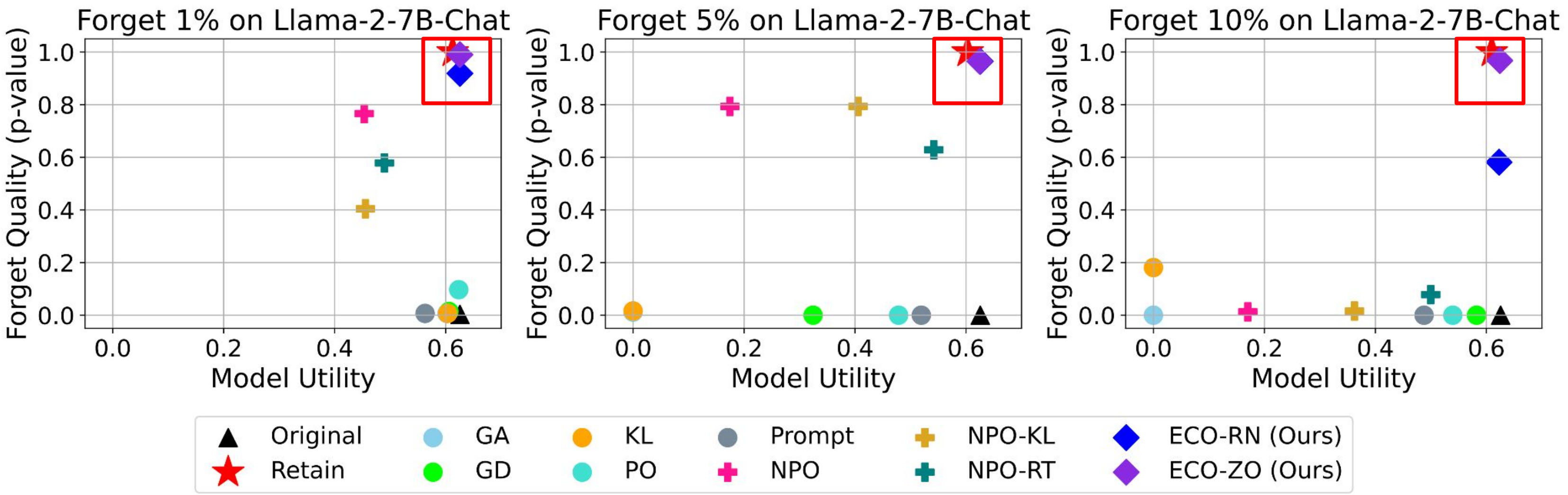
Copyrighted Unlearning on Book and News

The least different from the retained model

With minimal performance loss



Biography Unlearning on TOFU



Higher model utility and higher forget quality are better.

Conclusion

- A simple way to achieve unlearning for MaaS
- Scalable to any model size with no additional compute
- Ensure unlearning while largely preserve original performance on benign tasks

Model-as-a-service (MaaS) ✅

Future Work

- Why does it work?
    - (We provide a simple hypothesis in the appendix of paper.)
- Adapt to open-weight LLMs without the need of a classifier