Large Language Model Unlearning

Yuanshun Yao[1*] Xiaojun Xu[2] Yang Liu[3*]

[1]Meta GenAI

[2]ByteDance Research

[3]University of California, Santa Cruz

[*] Work done while at ByteDance research.
Correspond to kevinyao@meta.com

LLM Unlearning

- If an LLM learns unwanted misbehavior, unlearn or “forget” them with samples that represent those problematic behaviors

- Use case
  1. Removing harmful responses
  2. Erasing copyrighted contents learned in training data
  3. Reducing hallucinations
  4. Adapting to quick policy changes

Benefit of LLM Unlearning

1. Only requires negative samples $\rightarrow$ easy to collect by auto red teaming

2. Fast (cost is comparable to just LLM finetuning)

3. Efficient when you know which training samples cause misbehaviors

Method

$$\theta_{t+1} \leftarrow \theta_t - \underbrace{\epsilon_1 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{fgt}}}_{\text{Unlearn Harm}} - \underbrace{\epsilon_2 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{rdn}}}_{\text{Random Mismatch}} - \underbrace{\epsilon_3 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{nor}}}_{\text{Maintain Performance}}$$

Method

$$\theta_{t+1} \leftarrow \theta_t - \underbrace{\epsilon_1 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{fgt}}}_{\text{Unlearn Harm}} - \underbrace{\epsilon_2 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{rdn}}}_{\text{Random Mismatch}} - \underbrace{\epsilon_3 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{nor}}}_{\text{Maintain Performance}}$$

- $\nabla_{\theta_t} \mathcal{L}_{\text{fgt}} := - \sum_{(x^{\text{fgt}}, y^{\text{fgt}}) \in D^{\text{fgt}}} L(x^{\text{fgt}}, y^{\text{fgt}}; \theta_t)$
  - Gradient Ascent to forget

Method

$$\theta_{t+1} \leftarrow \theta_t - \underbrace{\epsilon_1 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{fgt}}}_{\text{Unlearn Harm}} - \underbrace{\epsilon_2 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{rdn}}}_{\text{Random Mismatch}} - \underbrace{\epsilon_3 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{nor}}}_{\text{Maintain Performance}}$$

- $\nabla_{\theta_t} \mathcal{L}_{\text{fgt}} := - \sum_{(x^{\text{fgt}}, y^{\text{fgt}}) \in D^{\text{fgt}}} L(x^{\text{fgt}}, y^{\text{fgt}}; \theta_t)$
  - Gradient Ascent to forget
- $\nabla_{\theta_t} \mathcal{L}_{\text{rdn}} := \sum_{(x^{\text{fgt}}, \cdot) \in D^{\text{fgt}}} \frac{1}{|\mathcal{Y}^{\text{rdn}}|} \sum_{y^{\text{rdn}} \in \mathcal{Y}^{\text{rdn}}} L(x^{\text{fgt}}, y^{\text{rdn}}; \theta_t)$
  - Forced the model to predict random answers unrelated to x^{fgt}
    - Help LLM forget unwanted outputs on x^{fgt}
  - We empirically find it helps preserve the normal utility with theoretical analysis

Method

$$\theta_{t+1} \leftarrow \theta_t - \underbrace{\epsilon_1 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{fgt}}}_{\text{Unlearn Harm}} - \underbrace{\epsilon_2 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{rdn}}}_{\text{Random Mismatch}} - \underbrace{\epsilon_3 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{nor}}}_{\text{Maintain Performance}}$$

- $\nabla_{\theta_t} \mathcal{L}_{\text{nor}} := \sum_{(x^{\text{nor}}, y^{\text{nor}}) \in D^{\text{nor}}} \sum_{i=1}^{|y^{\text{nor}}|} \text{KL}\left(h_{\theta^o}(x^{\text{nor}}, y^{\text{nor}}_{<i}) || h_{\theta_t}(x^{\text{nor}}, y^{\text{nor}}_{<i})\right)$
  - Forward KL (i.e. $KL(\theta_{ref}||\theta_t)$) rather than backward KL (i.e. $KL(\theta_t||\theta_{ref})$) in RLHF (i.e. sampling)

Method

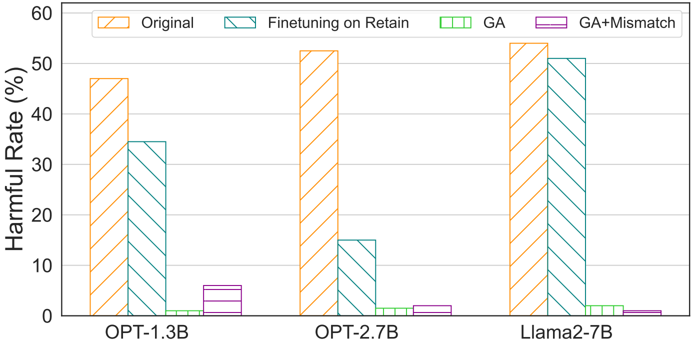
$$\theta_{t+1} \leftarrow \theta_t - \underbrace{\epsilon_1 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{fgt}}}_{\text{Unlearn Harm}} - \underbrace{\epsilon_2 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{rdn}}}_{\text{Random Mismatch}} - \underbrace{\epsilon_3 \cdot \nabla_{\theta_t} \mathcal{L}_{\text{nor}}}_{\text{Maintain Performance}}$$

- $\nabla_{\theta_t} \mathcal{L}_{\text{nor}} := \sum_{(x^{\text{nor}}, y^{\text{nor}}) \in D^{\text{nor}}} \sum_{i=1}^{|y^{\text{nor}}|} \text{KL}\left(h_{\theta^o}(x^{\text{nor}}, y^{\text{nor}}_{<i}) || h_{\theta_t}(x^{\text{nor}}, y^{\text{nor}}_{<i})\right)$
  - Forward KL (i.e. $KL(\theta_{ref}||\theta_t)$) rather than backward KL (i.e. $KL(\theta_t||\theta_{ref})$) in RLHF (i.e. sampling)
- All GA and GD are done on y (response) only rather than (x, y)

$$L(x, y; \theta) := \sum_{i=1}^{|y|} \ell\left(h_\theta(x, y_{<i}), y_i\right)$$

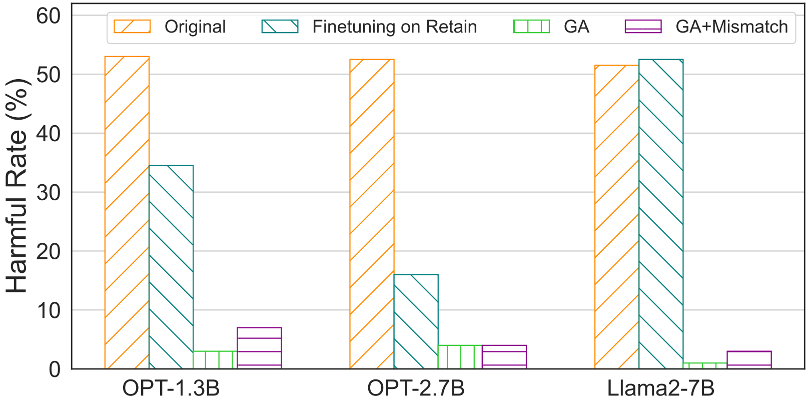
Application: Unlearning Harmfulness

- Forgetting data: PKU-SafeRLHF; Normal data: TruthfulQA



Forget unlearned samples

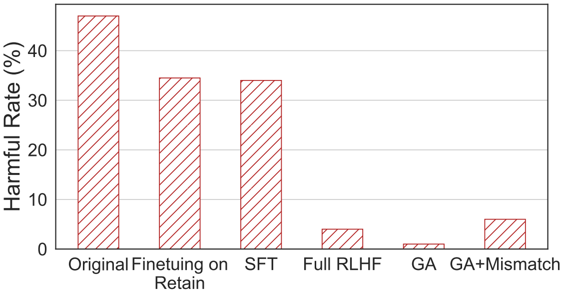
Outputs on Unseen Prompts



Generalized to unseen harmful prompts

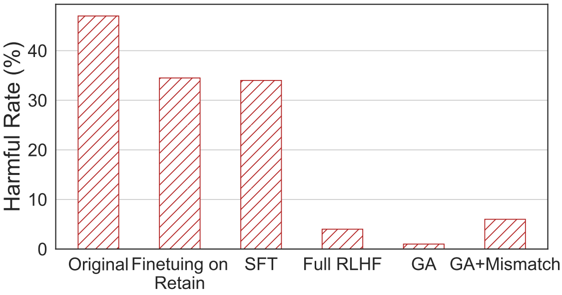
See the paper for the application of unlearning (1) copyrighted contents and (2) hallucinations

Ablation: Comparing to RLHF

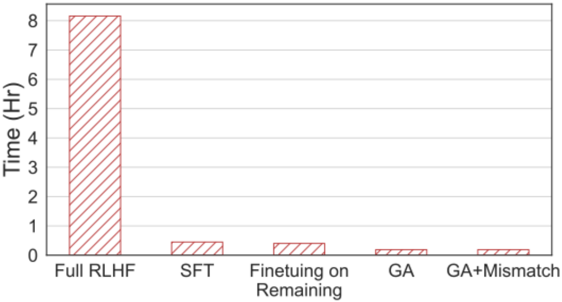


Comparable performance to RLHF

Ablation: Comparing to RLHF



Comparable performance to RLHF



Only 2% of time

Takeaways

- We should not conclude; this is a growing area [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

Takeaways

- We should not conclude; this is a growing area [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
- Targeted and direct unlearning could be an alternative in alignment

Thanks!

Email: kevinyao@meta.com

Full Paper

Code

References I

[1] Andrei Muresanu et al. "Unlearnable algorithms for in-context learning". In: *arXiv preprint arXiv:2402.00751* (2024).

[2] Shitong Duan et al. "Negating Negatives: Alignment without Human Positive Samples via Distributional Dispreference Optimization". In: *arXiv preprint arXiv:2403.03419* (2024).

[3] Sungmin Cha et al. "Towards Robust and Cost-Efficient Knowledge Unlearning for Large Language Models". In: *arXiv preprint arXiv:2408.06621* (2024).

[4] Weijia Shi et al. "Muse: Machine unlearning six-way evaluation for language models". In: *arXiv preprint arXiv:2407.06460* (2024).

[5] Ruiqi Zhang et al. "Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning". In: *arXiv preprint arXiv:2404.05868* (2024).

References II

[6] Zheyuan Liu et al. "Towards Safer Large Language Models through Machine Unlearning". In: *arXiv preprint arXiv:2402.10058* (2024).

[7] Chongyang Gao et al. "Practical unlearning for large language models". In: *arXiv preprint arXiv:2407.10223* (2024).

[8] James Y Huang et al. "Offset Unlearning for Large Language Models". In: *arXiv preprint arXiv:2404.11045* (2024).

[9] Zhexin Zhang et al. "Safe unlearning: A surprisingly effective and generalizable solution to defend against jailbreak attacks". In: *arXiv preprint arXiv:2407.02855* (2024).

[10] Sijia Liu et al. "Rethinking Machine Unlearning for Large Language Models". In: *arXiv preprint arXiv:2402.08787* (2024).