Data-Efficient Task Unlearning in Language Models

Andrzej Szablewski, Szymon Duchniewicz, Zhe Yu, Yadong Liu

University College London, Department of Computer Science

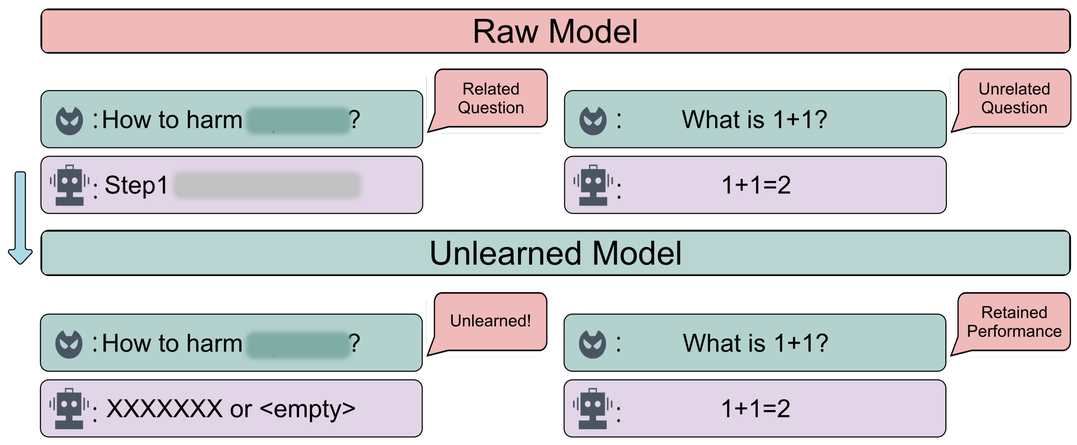
What is Task Unlearning in LMs?

Depending on the application, some of the abilities of Large Language Models may be undesirable. While model alignment is possible, it is easier to gather negative rather than positive prompt-response pairs.

Task unlearning has two goals:

1. **Decreasing model performance on an undesired task.**
2. **Keeping the remaining abilities of the model unaffected.**

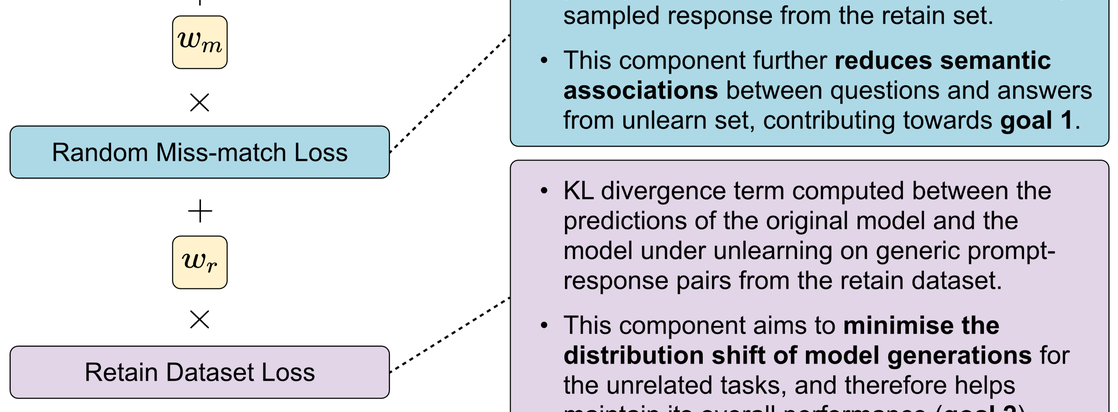
Application: Improving Model Safety



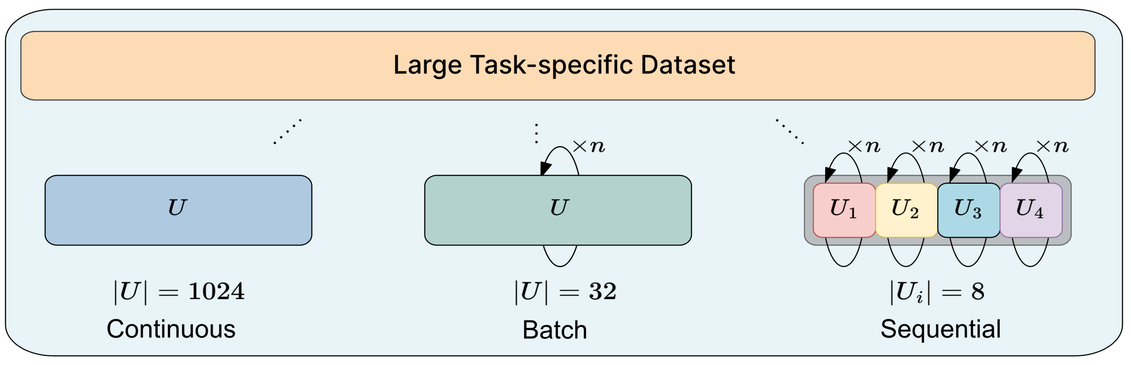
- Unlearn the OPT-1.3B model responding in a toxic or harmful manner.
- Measure model safety using a question-answering task *BeaverTails-Evaluation* and an expert model classifying harmful model generations.
- Monitor its remaining capabilities using a set of generic task-specific benchmarks (e.g. *Winogrande, MMLU, ARC*).

How to Unlearn a Particular Task?

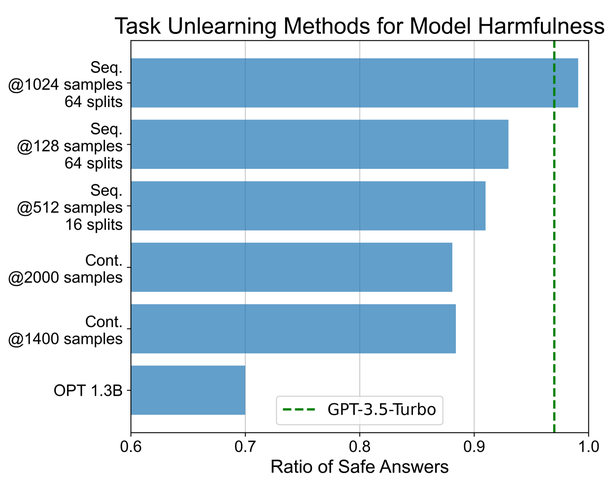
- Apply machine unlearning and optimise for the goals using a combination of undesired and generic samples!



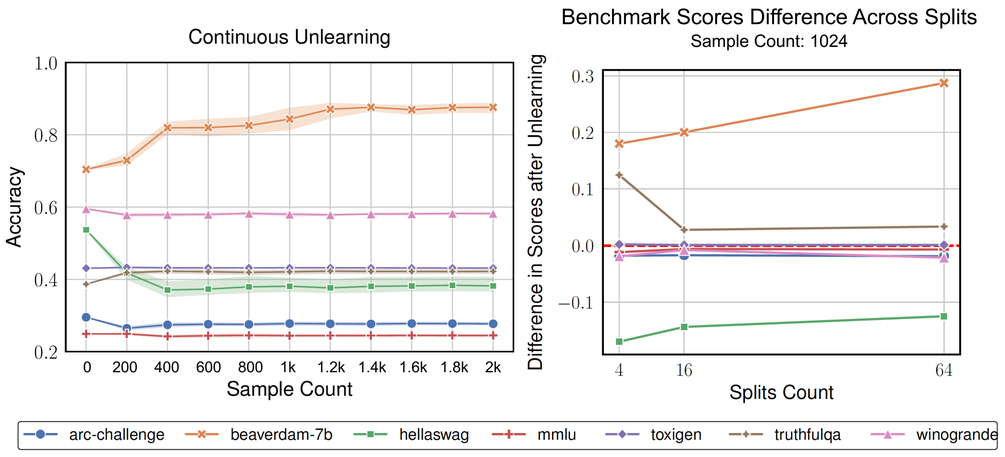
- In addition to gradient ascent on undesired samples, minimise unlearning of the remaining model capabilities.
- Split undesired samples into small batches and update model parameters sequentially after each batch.



Smaller Batches Better Preserve Overall Model Capabilities

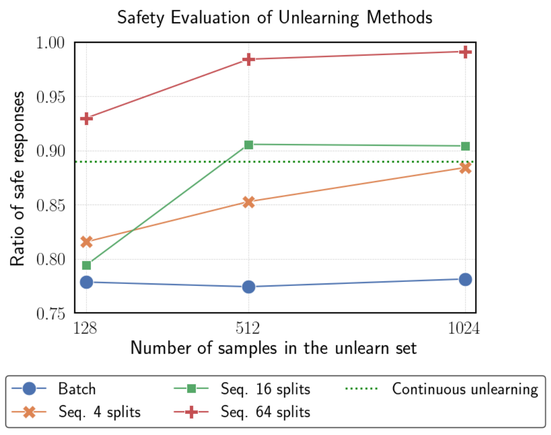


- Sequential parameter updates result in more stable and successful task unlearning.
- Harmfulness unlearning affects the natural language inference abilities of the model.
- Iterative unlearning of a small batch performs better than a single model update with a big batch and a scaled learning rate.

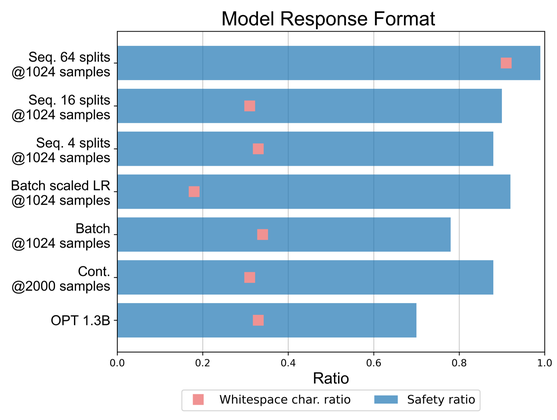


Data-Efficiency of Sequential Unlearning

- Increasing the number of splits leads to better unlearning performance in the sequential approach.
- Unlearning by computing gradients over larger batches requires a significant amount of data.



Change of Model Response Format



Unlearning Method	Avg. Model Response Length (characters)
OPT 1.3B (base model)	1268
Continuous (2000 samples)	72
Batch (1024 samples)	717
Batch (1024 samples, scaled lr)	639
Sequential (1024 samples, 4 splits)	85
Sequential (1024 samples, 16 splits)	461
Sequential (1024 samples, 64 splits)	606

- Retaining on *TruthfulQA* leads to overfitting and increased model performance on this task.
- All unlearning methods affect the model response format.