

LLM



Model loader

Architecture

1. Llama 2/3
2. Zephyr
3. Phi 1.5/3.5
4. Gemma

Load as

1. Full precision
2. Quantized
3. Probed model

Data worker

Data Processor

1. Pretraining dataset
2. Chat dataset
3. Completion dataset
4. Unlearning dataset (forget x retain)

Raw data



Trainer

Finetune Trainer

Unlearning Trainer

- | | |
|---------------|-----------|
| 1. GradAscent | 4. DPO |
| 2. GradDiff | 5. RMU |
| 3. NPO | 6. SimNPO |

Interventions

Stress Test via

1. Re-learning attacks
2. Quantization attack
3. Training a probe on a layer



Unlearned LLM



Stress Tested LLM

Really
unlearned?

Evaluator

Benchmarks

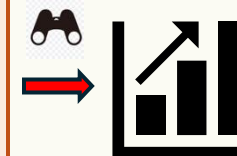
TOFU
MUSE
WMDP

Metrics

- | | |
|----------------|------------|
| 1. ROUGE | 4. EM |
| 2. Probability | 5. MIA |
| 3. ES | 6. LM Eval |



Unlearned
model
Evaluations



Stress tested
model
Evaluations