

Unlearning with Anonymized Data

30-05-2025

Update

- Last week:
 - Experimented with and demonstrated the performances of models unlearned using gradient-based methods on data anonymized with encryption (RSA), and NER followed by synthetic entity substitution.
- This week:
 - Experimented with LLM as an anonymizer
 - Exploring oblivious unlearning
 - Started unlearning with RL

LLM as an anonymizer

- Model: Llama 70B using Grok
- Prompt:

```
"role": "user",
"content":
f"""You are an advanced QA anonymization model trained to rewrite author-related questions and answers by removing all personally identifiable information (PII)
|while preserving meaning and answerability. Follow these rules strictly:

### Core Principles:
1. **Generalize, Don't Memorize**:
   - Never reuse phrasing from examples. Treat each input as unique.
   - Use context (era, genre, nationality, awards) to replace PII descriptively.

2. **Anonymization Rules**:
   - **Authors**: Replace names with descriptors (e.g., "a 20th-century Japanese novelist" for Haruki Murakami).
   - **Books**: Keep titles only if they don't reveal the author (e.g., "*1984*" → "a dystopian novel from the 1940s" if Orwell is named).
   - **Specifics**: Mask exact dates/locations (e.g., "born in 1965" → "born in the mid-1960s").

3. **Natural Language**:
   - Avoid robotic templates. Rewrite fluently (e.g., "This author..." → "They...").
   - If the question references uniqueness (e.g., "the only author to..."), generalize it (e.g., "a rare author who...").

4. **Output**:
   - Return **only** the anonymized question and answer in this format:
     Anonymized Question: [Your rewrite]
     Anonymized Answer: [Your rewrite]

### Input to Anonymize:
Original Question: {question}
Original Answer: {answer}

### Task:
Apply the above rules rigorously. Do not mirror examples—generate fresh anonymization every time."""

```

LLM as an anonymizer

- Sample outputs:
 - {"original_question": "What other topics has Hsiao Yun-Hwa written about apart from leadership?", "original_answer": "Apart from leadership, Hsiao Yun-Hwa has also discussed topics closely related to it, such as diversity, inclusion and team-building within her books.", "anonymized_question": "What other topics has a Taiwanese author known for leadership expertise written about?", "anonymized_answer": "Apart from leadership, this author has also explored closely related topics, such as diversity, inclusion, and team-building in their published works."}
 - {"original_question": "Has Jaime Vasquez earned any awards for his controversial works?", "original_answer": "Jaime Vasquez was bestowed with the prestigious Edgar Allan Poe Award for Best Fact Crime, which is an illustrious accolade in the domain of crime fiction and non-fiction. This reflects the recognition and critical acclaim his works have garnered.", "anonymized_question": "Has a provocative writer from the contemporary era earned any accolades for their works?", "anonymized_answer": "A renowned award in the realm of crime fiction and non-fiction was bestowed upon a celebrated author, reflecting the widespread recognition and critical acclaim their works have garnered."}

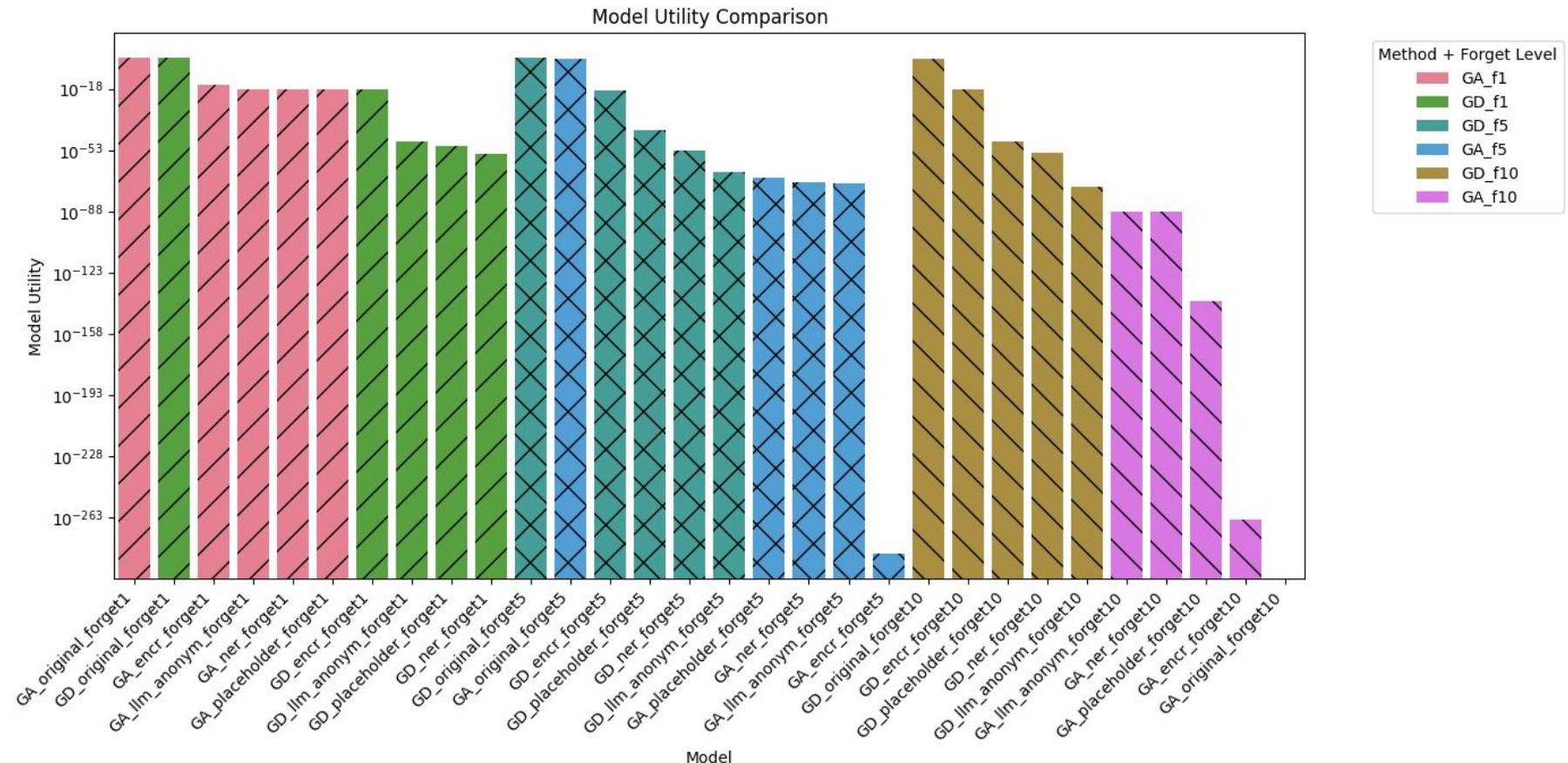
LLM as an anonymizer – results

LLM as anonymizer					
Phi-Retain set					
Model	Model utilit	Forget qualit	Prob.	ROUGE	Truth ratio
GA_llm_anonym_forget1_model	1.07E-18	0.0005039436	3.53E-19	0.7510473206	3.53E-19
GA_llm_anonym_forget5_model	1.35E-72	1.94E-119	3.00E-73	0.5626788108	3.00E-73
GA_llm_anonym_forget10_model	1.20E-88	1.06E-239	2.67E-89	0.4652703447	2.67E-89
GD_llm_anonym_forget1_model	1.04E-48	1.86E-23	2.32E-49	0.3127027198	2.32E-49
GD_llm_anonym_forget5_model	4.42E-66	1.94E-119	9.82E-67	0.1395884585	9.82E-67
GD_llm_anonym_forget10_model	1.03E-74	1.06E-239	2.29E-75	0.1783112535	2.29E-75
Phi-Forget set					
Model	Model utilit	Forget qualit	Prob.	ROUGE	Truth ratio
GA_llm_anonym_forget1_model	1.07E-18	0.0005039436	1.23E-27	0.7262456056	1.23E-27
GA_llm_anonym_forget5_model	1.35E-72	1.94E-119	4.99E-79	0.544924934	4.99E-79
GA_llm_anonym_forget10_model	1.20E-88	1.06E-239	3.23E-92	0.4626105143	3.23E-92
GD_llm_anonym_forget1_model	1.04E-48	1.86E-23	1.25E-78	0.3020991397	1.25E-78
GD_llm_anonym_forget5_model	4.42E-66	1.94E-119	1.13E-84	0.1410311267	1.13E-84
GD_llm_anonym_forget10_model	1.03E-74	1.06E-239	4.05E-74	0.1812039103	4.05E-74

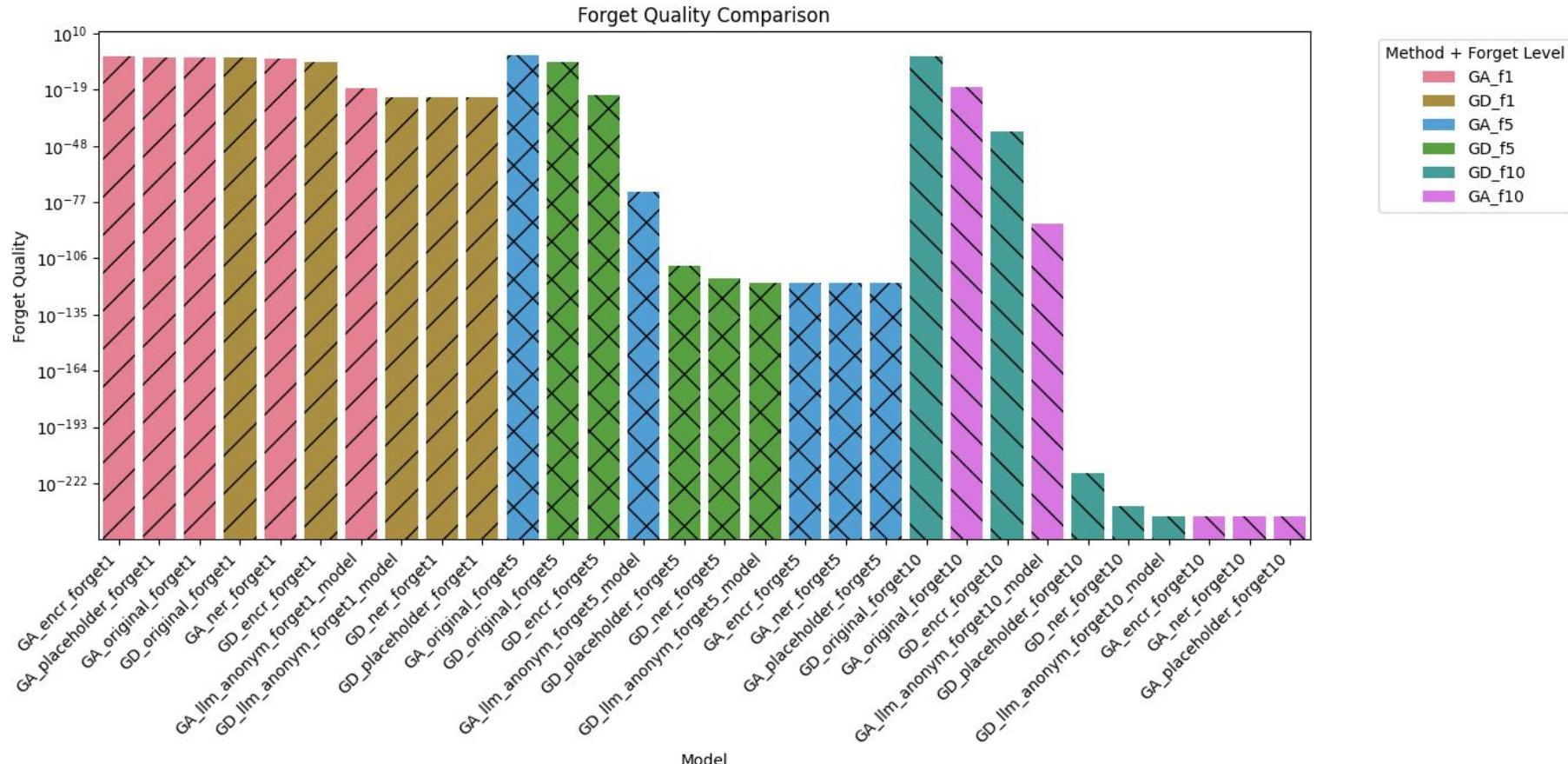
LLM as an anonymizer — results

Phi-Real authors					
Model	Model utilit	Forget qualit	Prob.	ROUGE	Truth ratio
GA_llm_anonym_forget1_model	1.07E-18	0.0005039436	7.33E-19	0.4106817777	7.33E-19
GA_llm_anonym_forget5_model	1.35E-72	1.94E-119	5.35E-67	0.2596937961	5.35E-67
GA_llm_anonym_forget10_model	1.20E-88	1.06E-239	3.65E-80	0.1721148239	3.65E-80
GD_llm_anonym_forget1_model	1.04E-48	1.86E-23	5.54E-45	0.1662608014	5.54E-45
GD_llm_anonym_forget5_model	4.42E-66	1.94E-119	1.39E-51	0.1289106124	1.39E-51
GD_llm_anonym_forget10_model	1.03E-74	1.06E-239	4.65E-55	0.1444658237	4.65E-55
Phi-Real world					
Model	Model utilit	Forget qualit	Prob.	ROUGE	Truth ratio
GA_llm_anonym_forget1_model	1.07E-18	0.0005039436	3.21E-16	0.4102605728	3.21E-16
GA_llm_anonym_forget5_model	1.35E-72	1.94E-119	4.04E-64	0.2863166909	4.04E-64
GA_llm_anonym_forget10_model	1.20E-88	1.06E-239	3.65E-75	0.2573105638	3.65E-75
GD_llm_anonym_forget1_model	1.04E-48	1.86E-23	1.61E-29	0.2336392065	1.61E-29
GD_llm_anonym_forget5_model	4.42E-66	1.94E-119	1.44E-34	0.1736077719	1.44E-34
GD_llm_anonym_forget10_model	1.03E-74	1.06E-239	5.68E-38	0.1757017551	5.68E-38

Model utility comparison



Forget quality comparison



Oblivious unlearning

- Oblivious Unlearning by Learning: Machine Unlearning Without Exposing Erased Data – Submitted to ICLR 2025 – Rejected – [Paper link](#)
- Core objective:
 - Enable privacy-preserving machine unlearning by preventing exposure of the forget set to the server.
 - Achieve unlearning without server-side modifications or retraining from scratch.
- Proposed approach: Oblivious Unlearning by Learning (OUbL)
 - Replace the forget set with synthetic noise-augmented samples.
 - Server performs standard incremental training on synthetic data, oblivious to unlearning intent.
- Pipeline overview
 - User: Synthesize noisy samples from forget set.
 - Server: Train on synthetic data → implicitly unlearns forget set.
 - Output: Model behaves as if forget set was never used.

Oblivious unlearning

- Noisy sample generation technique: Gradient-matching noise synthesis.
 - Compute gradients of the forget set w.r.t. model loss.
 - Generate noise to approximate the negative influence of forget set gradients.
 - Blend noise into auxiliary samples (e.g., random/public data) to mask original data.
- Datasets used: MNIST, CIFAR10, CelebA.
- Models used: 5-layer multi-layer perceptron (MLP) connected by ReLU on MNIST, a ResNet18 on CIFAR10 and a 7-layer convolutional neural network (CNN) on CelebA.
- Unlearning methods used: SISA, VBU (approximate unlearning), BFU (federated unlearning), HBFU (federated unlearning).
- Metrics used: Reconstruction similarity, Backdoor accuracy, Unlearning update similarity, Model accuracy, Running time.

Next steps

- From the various experiments conducted, it is evident that the traditional gradient-based unlearning algorithms drastically underperform on anonymized datasets.
- Experiments using non-gradient based unlearning methods is in progress.
 - A baseline experiment using RL for anonymized unlearning is in progress.
 - Exploration and shortlisting of potential non-gradient based methods that are generalizable but not dataset-specific is ongoing.
- Focus on designing an unlearning algorithm:
 - A solution in the space of meta learning + sensitivity suppression + oblivious unlearning is being explored.

Unlearning with Anonymized Data

06-06-2025 & 13-06-2025

Update

- Completed/ongoing:
 - Family of models on TOFU benchmark
 - T-test for performance scores
 - Design an unlearning algorithm
 - RL based unlearning
 - Key features identification with LIME/SHAP
- Yet to start:
 - LLM as an anonymizer – prompt change to control the extreme level of anonymization
 - Explore other generalizable non-gradient based unlearning methods
 - Additional evaluation metrics to be added

Family of models on TOFU benchmark – Deepseek_Qwen2.5_1.5B

Encryption-anonymization					
Deepseek_Qwen2.5_1.5B-Retain set					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_encr_forget1	1.12E-78	1.53E-18	2.49E-79	0.266409991	2.49E-79
GA_encr_forget5	4.81E-02	1.94E-119	0.00E+00	0.0310757489	0.00E+00
GA_encr_forget10	2.55E-02	1.94E-119	0	0.0141347403	0
GD_encr_forget1	6.12E-77	2.94E-17	1.36E-77	0.2757764213	1.36E-77
GD_encr_forget5	1.42E-175	1.94E-119	3.54E-176	0.0050084997	3.54E-176
GD_encr_forget10	9.15E-175	1.94E-119	2.61E-175	0.0002168458	2.61E-175
Deepseek_Qwen2.5_1.5B-Forget set					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_encr_forget1	1.12E-78	1.53E-18	1.34E-109	0.2688548463	1.34E-109
GA_encr_forget5	4.81E-02	1.94E-119	0	0.0295829251	0
GA_encr_forget10	2.55E-02	1.94E-119	0	0.0132726213	0
GD_encr_forget1	6.12E-77	2.94E-17	6.67E-111	0.2792718678	6.67E-111
GD_encr_forget5	1.42E-175	1.94E-119	6.20E-202	0.0057632226	6.20E-202
GD_encr_forget10	9.15E-175	1.94E-119	2.81E-206	0.0004280099	2.81E-206

Family of models on TOFU benchmark – Deepseek_Qwen2.5_1.5B

Deepseek_Qwen2.5_1.5B-Real authors					
Model	Model utility	Forget quality Prob.	ROUGE	Truth ratio	
GA_encr_forget1	1.12E-78	1.53E-18	4.11E-66	0.0993817625	4.11E-66
GA_encr_forget5	4.81E-02	1.94E-119	0.00E+00	0.0500456077	0.00E+00
GA_encr_forget10	2.55E-02	1.94E-119	0.00E+00	0.0294363989	0.00E+00
GD_encr_forget1	6.12E-77	2.94E-17	2.89E-65	0.1082369903	2.89E-65
GD_encr_forget5	1.42E-175	1.94E-119	1.54E-129	0.0013333333	1.54E-129
GD_encr_forget10	9.15E-175	1.94E-119	3.81E-132	0	3.81E-132
Deepseek_Qwen2.5_1.5B-Real world					
Model	Model utility	Forget quality Prob.	ROUGE	Truth ratio	
GA_encr_forget1	1.12E-78	1.53E-18	9.52E-55	0.1532737027	9.52E-55
GA_encr_forget5	4.81E-02	1.94E-119	0.00E+00	0.0983585699	0.00E+00
GA_encr_forget10	2.55E-02	1.94E-119	0.00E+00	0.0776085627	0.00E+00
GD_encr_forget1	6.12E-77	2.94E-17	1.24E-59	0.1500257752	1.24E-59
GD_encr_forget5	1.42E-175	1.94E-119	1.42E-109	0	1.42E-109
GD_encr_forget10	9.15E-175	1.94E-119	2.11E-107	0	2.11E-107

Family of models on TOFU benchmark – Deepseek_Qwen2.5_1.5B

NER-anonymization					
Deepseek_Qwen2.5_1.5B-Retain set					
Model	Model utility	Forget quality Prob.	ROUGE	Truth ratio	
GA_ner_forget1_deepseek_qwen2.1	1.12E-78	1.53E-18	2.25E-39	0.3762363053	2.25E-39
GA_ner_forget5_deepseek_qwen2.1	0.04813239146	1.94E-119	2.04E-119	0.3427670865	2.04E-119
GA_ner_forget10_deepseek_qwen2	0.02550925033	1.94E-119	0	0.0033483555	0
Deepseek_Qwen2.5_1.5B-Forget set					
Model	Model utility	Forget quality Prob.	ROUGE	Truth ratio	
GA_ner_forget1_deepseek_qwen2.1	1.12E-78	1.53E-18	5.79E-56	0.3860339592	5.79E-56
GA_ner_forget5_deepseek_qwen2.1	0.04813239146	1.94E-119	1.61E-138	0.3522481348	1.61E-138
GA_ner_forget10_deepseek_qwen2	0.02550925033	1.94E-119	0	0.0049382368	0
Deepseek_Qwen2.5_1.5B-Real authors					
Model	Model utility	Forget quality Prob.	ROUGE	Truth ratio	
GA_ner_forget1_deepseek_qwen2.1	1.12E-78	1.53E-18	2.67E-26	0.3433876083	2.67E-26
GA_ner_forget5_deepseek_qwen2.1	0.04813239146	1.94E-119	4.76E-103	0.2090656911	4.76E-103
GA_ner_forget10_deepseek_qwen2	0.02550925033	1.94E-119	0	0	0
Deepseek_Qwen2.5_1.5B-Real world					
Model	Model utility	Forget quality Prob.	ROUGE	Truth ratio	
GA_ner_forget1_deepseek_qwen2.1	1.12E-78	1.53E-18	6.67E-21	0.3839403185	6.67E-21
GA_ner_forget5_deepseek_qwen2.1	0.04813239146	1.94E-119	1.01E-86	0.2602058136	1.01E-86
GA_ner_forget10_deepseek_qwen2	0.02550925033	1.94E-119	0	0	0

Family of models on TOFU benchmark – Deepseek_Qwen2.5_1.5B

Placeholder-anonymization					
Deepseek_Qwen2.5_1.5B-Retain set					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_forget1_deepseek_qwen2.5_1.!	3.74E-39	1	8.31E-40	0.3755686131	8.31E-40
GA_forget5_deepseek_qwen2.5_1.!	2.31E-87	3.36E-63	5.14E-88	0.3710662038	5.14E-88
GA_forget10_deepseek_qwen2.5_1	0.01696544783	1.94E-119	0	0.0169654478	0
Deepseek_Qwen2.5_1.5B-Forget set					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_forget1_deepseek_qwen2.5_1.!	3.74E-39	1	9.54E-56	0.3866683535	9.54E-56
GA_forget5_deepseek_qwen2.5_1.!	2.31E-87	3.36E-63	3.15E-97	0.374379255	3.15E-97
GA_forget10_deepseek_qwen2.5_1	0.01696544783	1.94E-119	0	0.0227872933	0
Deepseek_Qwen2.5_1.5B-Real authors					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_forget1_deepseek_qwen2.5_1.!	3.74E-39	1	2.16E-26	0.341209486	2.16E-26
GA_forget5_deepseek_qwen2.5_1.!	2.31E-87	3.36E-63	3.58E-69	0.2788141974	3.58E-69
GA_forget10_deepseek_qwen2.5_1	0.01696544783	1.94E-119	0	0	0
Deepseek_Qwen2.5_1.5B-Real world					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_forget1_deepseek_qwen2.5_1.!	3.74E-39	1	6.71E-21	0.3813435514	6.71E-21
GA_forget5_deepseek_qwen2.5_1.!	2.31E-87	3.36E-63	7.72E-62	0.3326366743	7.72E-62
GA_forget10_deepseek_qwen2.5_1	0.01696544783	1.94E-119	0	0	0

Family of models on TOFU benchmark – Gemma2_2B

Encryption-anonymization					
Gemma2_2B-Retain set					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_encr_forget1	1.15E-15	0.0705313	2.40E-15	3.316788	1.11E-14
GA_encr_forget5	3.30E-283	1.28E-118	1.16E-283	0.039611	4.26E-284
GA_encr_forget10	1.70E-263	1.14E-238	0	0.106441	0
GD_encr_forget1	5.10E-18	0.0001184	4.81E-19	2.312957	1.09E-18
GD_encr_forget5	1.70E-18	7.47E-22	2.05E-19	4.196698	2.73E-19
GD_encr_forget10	4.35E-18	5.01E-40	4.76E-19	1.840655	1.22E-18
Gemma2_2B-Forget set					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_encr_forget1	1.15E-15	0.0705313	4.49E-22	3.42571	3.19E-22
GA_encr_forget5	3.30E-283	1.28E-118	0	0.099338	0
GA_encr_forget10	1.70E-263	1.14E-238	0	0.028117	0
GD_encr_forget1	5.29E-18	0.0000882	7.66E-27	2.478894	7.15E-27
GD_encr_forget5	6.08E-19	8.30E-22	7.18E-24	8.142418	4.06E-23
GD_encr_forget10	1.88E-18	1.99E-40	1.27E-20	2.621656	7.11E-21

Family of models on TOFU benchmark – Gemma2_2B

Gemma2_2B-Real authors					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_encr_forget1	1.15E-15	0.0705313	5.02E-16	3.049806	2.92E-16
GA_encr_forget5	3.30E-283	1.28E-118	1.11E-172	0.670533	9.07E-173
GA_encr_forget10	1.70E-263	1.14E-238	3.60E-264	0.036135	1.97E-264
GD_encr_forget1	2.43E-18	0.0001942	1.48E-15	4.542497	7.06E-16
GD_encr_forget5	7.29E-19	1.44E-21	1.40E-15	1.896053	3.24E-15
GD_encr_forget10	4.24E-18	3.08E-40	2.36E-15	4.356409	2.23E-15
Gemma2_2B-Real world					
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio
GA_encr_forget1	1.15E-15	0.0705313	1.09E-11	4.549604	1.10E-11
GA_encr_forget5	3.30E-283	1.28E-118	6.02E-57	1.554084	3.62E-56
GA_encr_forget10	1.70E-263	1.14E-238	1.77E-105	0.947281	8.42E-106
GD_encr_forget1	3.52E-18	0.0000512	5.32E-13	3.17449	6.75E-13
GD_encr_forget5	3.25E-19	1.17E-22	6.50E-13	3.207812	9.36E-13
GD_encr_forget10	3.71E-18	2.66E-40	5.89E-13	1.375159	6.13E-13

Non-Gradient based Unlearning methods

- SKU – Selective Knowledge Negation Unlearning
 - Remove harmful knowledge by directly editing model parameters by neutralizing learned directions.
 - Identify “knowledge vectors” linked to harmful info.
 - Negate and subtract these vectors from model weights.
 - No retraining or backpropagation — purely parameter-level edits.
 - *“Towards Safer LLMs through Machine Unlearning” (2024) – [arXiv:2402.10058](#)*
- UnSTAR – Unlearning with Self-Taught Anti-Sample Reasoning
 - Induce forgetting through targeted counterexamples.
 - Generate anti-samples with misleading or contradictory reasoning.
 - Feed them into the model to shift internal beliefs away from target knowledge.
 - No gradient updates — unlearning driven by exposure, not optimization.
 - *“UnSTAR: Self-Taught Anti-Sample Reasoning for LLMs” (2025 - ICLR rejected) – [arXiv:2410.17050](#)*
- Conditional Sparse Autoencoder Clamping
 - Suppress internal activations related to unwanted knowledge.
 - Train Sparse Autoencoder on model activations to extract interpretable features.
 - Identify and clamp (disable or fix) activations tied to harmful knowledge.
 - Does not require modifying base model weights.
 - *“Conditional SAE Clamping Works for Unlearning” (2025) – [arXiv:2503.11127](#)*

Non-Gradient based Unlearning methods

- UnSTAR – Unlearning with Self-Taught Anti-Sample Reasoning
 - Induce forgetting through targeted counterexamples.
 - Generate anti-samples with misleading or contradictory reasoning.
 - Feed them into the model to shift internal beliefs away from target knowledge.
 - No gradient updates — unlearning driven by exposure, not optimization.
 - “*UnSTAR: Self-Taught Anti-Sample Reasoning for LLMs*” (2025 - ICLR rejected) – [arXiv:2410.17050](https://arxiv.org/abs/2410.17050)
- On a lookout for other methods – in progress.

Unlearning method - Thought

- What if we don't forget the data – we forget the ability to use it?
- A small-scale controller network will be trained through meta-learning to yield gating vectors.
 - Anonymized samples are clustered into several data groups and each initialized with a gating vector.
 - The corresponding vectors are updated iteratively as different cluster groups are passed to the network.
 - The cluster groups are rearranged to form new groups upon an optimized gating vectors are reached.
 - The process starts with these optimized gating vectors as the initial states.
 - This repeats for a few times with different organizations of the cluster groups, meta tasks, or data samples.
- The resultant gating vectors are applied to the target model to scale or suppress internal pathways.
- The controller network may be used to dynamically generate gate vectors for each anonymized input.
- Meta objective:
 - Generalize unlearning behavior across multiple anonymized tasks or samples.
 - Learns to predict which internal pathways to inhibit based solely on anonymized inputs.
 - Enhance performance on retain set samples + impair performance on forget set samples + gate sparsity regularization

Takeaways

- There are quite a few interesting directions in the addressal of unlearning, particularly selective and concept unlearning.
 - Non-gradient based unlearning methods is a potential space to explore.
 - Thorough literature study and in-depth understanding of these methods is vital.
- Need to refine the proposed method with clear differentiating factors, and concrete pipeline.

Unlearning with Anonymized Data

27-06-2025 & 04-07-2025

Update

- Dataset created using RUPTA method - Robust utility preserving text anonymization
 - Trained and evaluated Phi-1.5B model on the anonymized data
 - Unlearning methods used: Gradient ascent, Gradient difference
- Implemented GUARD - Guided unlearning paper to check if gradients are really the problem!
- Explored literature – AlphaEdit, an approach for model editing is on cards.
- Designed a new approach

Evaluation - Privacy-Utility LLM-based anonymization (RUPTA)

P-U LLM as anonymizer					LLM as anonymizer						
On Forget5					Phi-Retain set						
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio	Model	Model uti	Forget quality	Prob.	ROUGE	Truth ratio
GA_P_U_LLM	2.69E-71	1.86E-23	6.02E-72	0.5522603534	6.02E-72	GA_llm_a	1.35E-72	1.94E-119	3.00E-73	0.5626788	3.00E-73
GD_P_U_LLM	4.77E-58	1.94E-119	1.06E-58	0.1156769821	1.06E-58	GD_llm_a	4.42E-66	1.94E-119	9.82E-67	0.1395884	9.82E-67
Phi-Forget set					Phi-Forget set						
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio	Model	Model uti	Forget quality	Prob.	ROUGE	Truth ratio
GA_P_U_LLM	2.69E-71	1.86E-23	1.28E-89	0.5395581019	1.28E-89	GA_llm_a	1.35E-72	1.94E-119	4.99E-79	0.5449249	4.99E-79
GD_P_U_LLM	4.77E-58	1.94E-119	1.49E-75	0.1187913368	1.49E-75	GD_llm_a	4.42E-66	1.94E-119	1.13E-84	0.1410311	1.13E-84
Phi-Real authors					Phi-Real authors						
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio	Model	Model uti	Forget quality	Prob.	ROUGE	Truth ratio
GA_P_U_LLM	2.69E-71	1.86E-23	9.80E-70	0.2663497246	9.80E-70	GA_llm_a	1.35E-72	1.94E-119	5.35E-67	0.2596937	5.35E-67
GD_P_U_LLM	4.77E-58	1.94E-119	2.26E-46	0.1320715832	2.26E-46	GD_llm_a	4.42E-66	1.94E-119	1.39E-51	0.1289106	1.39E-51
Phi-Real world					Phi-Real world						
Model	Model utility	Forget quality	Prob.	ROUGE	Truth ratio	Model	Model uti	Forget quality	Prob.	ROUGE	Truth ratio
GA_P_U_LLM	2.69E-71	1.86E-23	4.40E-65	0.2980789489	4.40E-65	GA_llm_a	1.35E-72	1.94E-119	4.04E-64	0.2863166	4.04E-64
GD_P_U_LLM	4.77E-58	1.94E-119	2.68E-32	0.1078479179	2.68E-32	GD_llm_a	4.42E-66	1.94E-119	1.44E-34	0.1736077	1.44E-34

GUARD: Guided Unlearning and Retention via Data Attribution for Large Language Models

- Targeted unlearning of specific facts/samples while preserving utility, using data attribution to identify influential training points.
- **Finetune LLM:** Initially finetune the LLM on the complete dataset (D_0) to establish a baseline.
- **Compute Proxy Attribution:** Calculate a **lightweight proxy attribution score** (ai) for each forget sample by taking the inner product of its gradient with the average gradient of retain samples, indicating its alignment with retained knowledge.
- **Derive Unlearning Weights:** Convert attribution scores into **non-uniform unlearning weights** (w_i) using a softmax function with temperature T , inversely proportional to ai (higher alignment = lower weight).
 - T controls the "softness" or "sharpness" of the resulting unlearning weights.
- **Weighted Unlearning:** Apply these computed weights during the unlearning process (e.g., gradient ascent), **selectively reducing the unlearning emphasis** on samples critical for retention.
- **Mitigate Degradation:** This weighted approach ensures **effective forgetting** while significantly **reducing unintended knowledge degradation** on the retain set.
- **Enhanced Utility:** Ultimately, GUARD guarantees **improved utility preservation** compared to standard unlearning methods, maintaining the model's overall performance.

GUARD on Placeholder anonymization, 10% FS - T=0.1

```
Proxy scores computed for 320 forget samples.  
Proxy scores:  
tensor([0.8501, 0.8501, 0.8474, 0.8474, 0.8622, 0.8622, 0.8060, 0.8060, 0.8401,  
       0.8401, 0.7445, 0.7445, 0.8210, 0.8210, 0.8655, 0.8655, 0.8456, 0.8456,  
       0.7217, 0.7217, 0.8180, 0.8180, 0.7257, 0.7257, 0.8528, 0.8528, 0.8604,  
       0.8604, 0.8370, 0.8370, 0.8467, 0.8467, 0.8272, 0.8272, 0.8672, 0.8672,  
       0.8354, 0.8354, 0.8398, 0.8398, 0.8389, 0.8389, 0.7549, 0.7549, 0.8502,  
       0.8502, 0.8147, 0.8147, 0.7921, 0.7921, 0.8353, 0.8353, 0.8141, 0.8141,  
       0.7935, 0.7935, 0.8544, 0.8544, 0.8330, 0.8330, 0.8403, 0.8403, 0.8423,  
       0.8423, 0.8367, 0.8367, 0.8405, 0.8405, 0.8688, 0.8688, 0.8371, 0.8371,  
       0.7886, 0.7886, 0.8206, 0.8206, 0.7895, 0.7895, 0.8279, 0.8279, 0.8527,  
       0.8527, 0.8095, 0.8095, 0.8693, 0.8693, 0.8309, 0.8309, 0.8307, 0.8307,  
       0.8276, 0.8276, 0.8213, 0.8213, 0.7972, 0.7972, 0.7401, 0.7401, 0.8552,  
       0.8552, 0.7438, 0.7438, 0.7804, 0.7804, 0.8106, 0.8106, 0.8437, 0.8437,  
       0.8344, 0.8344, 0.8118, 0.8118, 0.8546, 0.8546, 0.8017, 0.8017, 0.8175,  
       0.8175, 0.8191, 0.8191, 0.8342, 0.8342, 0.8270, 0.8270, 0.8611, 0.8611,  
       0.8596, 0.8596, 0.8344, 0.8344, 0.8041, 0.8041, 0.8305, 0.8305, 0.8187,  
       0.8187, 0.8394, 0.8394, 0.7134, 0.7134, 0.7918, 0.7918, 0.7789, 0.7789,  
       0.8522, 0.8522, 0.7581, 0.7581, 0.8498, 0.8498, 0.8264, 0.8264, 0.8647,  
       0.8647, 0.7720, 0.7720, 0.7474, 0.7474, 0.8322, 0.8322, 0.8637, 0.8637,  
       0.8671, 0.8671, 0.8027, 0.8027, 0.7236, 0.7236, 0.8352, 0.8352, 0.8414,  
       0.8414, 0.8288, 0.8288, 0.7945, 0.7945, 0.8049, 0.8049, 0.8688, 0.8688,  
       0.7930, 0.7930, 0.7346, 0.7346, 0.8550, 0.8550, 0.8788, 0.8788, 0.7892,  
       0.7892, 0.8282, 0.8282, 0.7590, 0.7590, 0.8029, 0.8029, 0.8209, 0.8209,  
       0.7266, 0.7266, 0.8465, 0.8465, 0.8113, 0.8113, 0.8652, 0.8652, 0.8616,  
       0.8616, 0.8475, 0.8475, 0.8404, 0.8404, 0.8868, 0.8868, 0.7849, 0.7849,  
       0.8461, 0.8461, 0.7896, 0.7896, 0.8094, 0.8094, 0.7656, 0.7656, 0.6886,  
       0.6886, 0.8616, 0.8616, 0.8440, 0.8440, 0.7963, 0.7963, 0.8392, 0.8392,  
       0.8345, 0.8345, 0.7833, 0.7833, 0.8039, 0.8039, 0.8304, 0.8304, 0.8172,  
       0.8172, 0.8343, 0.8343, 0.7865, 0.7865, 0.7754, 0.7754, 0.8417, 0.8417,  
       0.8264, 0.8264, 0.8360, 0.8360, 0.7724, 0.7724, 0.8535, 0.8535, 0.8081,  
       0.8081, 0.7650, 0.7650, 0.8455, 0.8455, 0.8646, 0.8646, 0.7481, 0.7481,  
       0.8478, 0.8478, 0.7540, 0.7540, 0.8471, 0.8471, 0.8080, 0.8080, 0.7316,  
       0.7316, 0.8759, 0.8759, 0.8201, 0.8201, 0.8780, 0.8780, 0.7565, 0.7565,  
       0.8616, 0.8616, 0.7859, 0.7859, 0.8506, 0.8506, 0.8412, 0.8412, 0.8442,  
       0.8442, 0.8190, 0.8190, 0.7940, 0.7940, 0.8117, 0.8117, 0.7852, 0.7852,  
       0.8231, 0.8231, 0.8614, 0.8614, 0.8490, 0.8490, 0.8060, 0.8060, 0.8390,  
       0.8390, 0.8375, 0.8375, 0.8286, 0.8286], device='cuda:0')
```

Weights:

```
tensor([0.0021, 0.0021, 0.0022, 0.0022, 0.0019, 0.0019, 0.0033, 0.0033, 0.0023,
       0.0023, 0.0060, 0.0060, 0.0028, 0.0028, 0.0018, 0.0018, 0.0022, 0.0022,
       0.0076, 0.0076, 0.0029, 0.0029, 0.0073, 0.0073, 0.0020, 0.0020, 0.0019,
       0.0019, 0.0024, 0.0024, 0.0022, 0.0022, 0.0026, 0.0026, 0.0018, 0.0018,
       0.0024, 0.0024, 0.0023, 0.0023, 0.0023, 0.0023, 0.0054, 0.0054, 0.0021,
       0.0021, 0.0030, 0.0030, 0.0038, 0.0038, 0.0024, 0.0024, 0.0030, 0.0030,
       0.0037, 0.0037, 0.0020, 0.0020, 0.0025, 0.0025, 0.0023, 0.0023, 0.0023,
       0.0023, 0.0024, 0.0024, 0.0023, 0.0023, 0.0017, 0.0017, 0.0024, 0.0024,
       0.0039, 0.0039, 0.0028, 0.0028, 0.0038, 0.0038, 0.0026, 0.0026, 0.0020,
       0.0020, 0.0031, 0.0031, 0.0017, 0.0017, 0.0025, 0.0025, 0.0025, 0.0025,
       0.0026, 0.0026, 0.0028, 0.0028, 0.0036, 0.0036, 0.0063, 0.0063, 0.0020,
       0.0020, 0.0061, 0.0061, 0.0042, 0.0042, 0.0031, 0.0031, 0.0022, 0.0022,
       0.0025, 0.0025, 0.0031, 0.0031, 0.0020, 0.0020, 0.0034, 0.0034, 0.0029,
       0.0029, 0.0029, 0.0029, 0.0025, 0.0025, 0.0026, 0.0026, 0.0019, 0.0019,
       0.0019, 0.0019, 0.0025, 0.0025, 0.0033, 0.0033, 0.0026, 0.0026, 0.0029,
       0.0029, 0.0023, 0.0023, 0.0082, 0.0082, 0.0038, 0.0038, 0.0043, 0.0043,
       0.0021, 0.0021, 0.0053, 0.0053, 0.0021, 0.0021, 0.0027, 0.0027, 0.0018,
       0.0018, 0.0046, 0.0046, 0.0059, 0.0059, 0.0025, 0.0025, 0.0018, 0.0018,
       0.0018, 0.0018, 0.0034, 0.0034, 0.0074, 0.0074, 0.0024, 0.0024, 0.0023,
       0.0023, 0.0026, 0.0026, 0.0037, 0.0037, 0.0033, 0.0033, 0.0017, 0.0017,
       0.0037, 0.0037, 0.0067, 0.0067, 0.0020, 0.0020, 0.0016, 0.0016, 0.0039,
       0.0039, 0.0026, 0.0026, 0.0052, 0.0052, 0.0034, 0.0034, 0.0028, 0.0028,
       0.0072, 0.0072, 0.0022, 0.0022, 0.0031, 0.0031, 0.0018, 0.0018, 0.0019,
       0.0019, 0.0022, 0.0022, 0.0023, 0.0023, 0.0015, 0.0015, 0.0040, 0.0040,
       0.0022, 0.0022, 0.0038, 0.0038, 0.0032, 0.0032, 0.0049, 0.0049, 0.0106,
       0.0106, 0.0019, 0.0019, 0.0022, 0.0022, 0.0036, 0.0036, 0.0023, 0.0023,
       0.0025, 0.0025, 0.0041, 0.0041, 0.0033, 0.0033, 0.0026, 0.0026, 0.0029,
       0.0029, 0.0025, 0.0025, 0.0040, 0.0040, 0.0044, 0.0044, 0.0023, 0.0023,
       0.0027, 0.0027, 0.0024, 0.0024, 0.0046, 0.0046, 0.0020, 0.0020, 0.0032,
       0.0032, 0.0049, 0.0049, 0.0022, 0.0022, 0.0018, 0.0018, 0.0058, 0.0058,
       0.0021, 0.0021, 0.0055, 0.0055, 0.0022, 0.0022, 0.0032, 0.0032, 0.0069,
       0.0069, 0.0016, 0.0016, 0.0028, 0.0028, 0.0016, 0.0016, 0.0054, 0.0054,
       0.0019, 0.0019, 0.0040, 0.0040, 0.0021, 0.0021, 0.0023, 0.0023, 0.0022,
       0.0022, 0.0029, 0.0029, 0.0037, 0.0037, 0.0031, 0.0031, 0.0040, 0.0040,
       0.0028, 0.0028, 0.0019, 0.0019, 0.0021, 0.0021, 0.0033, 0.0033, 0.0023,
       0.0023, 0.0024, 0.0024, 0.0026, 0.0026], device='cuda:0')
```

GUARD on Placeholder anonymization, 10% FS - T=3.5

```
Proxy scores computed for 320 forget samples.  
Proxy scores:  
tensor([0.8632, 0.8632, 0.8142, 0.8142, 0.8057, 0.8057, 0.8546, 0.8546, 0.8145,  
       0.8145, 0.7182, 0.7182, 0.8670, 0.8670, 0.8308, 0.8308, 0.8189, 0.8189,  
       0.8665, 0.8665, 0.8087, 0.8087, 0.8190, 0.8190, 0.8397, 0.8397, 0.8167,  
       0.8167, 0.7953, 0.7953, 0.8342, 0.8342, 0.7827, 0.7827, 0.8180, 0.8180,  
       0.7500, 0.7500, 0.8417, 0.8417, 0.8093, 0.8093, 0.8676, 0.8676, 0.8307,  
       0.8307, 0.8499, 0.8499, 0.8433, 0.8433, 0.8564, 0.8564, 0.8421, 0.8421,  
       0.8254, 0.8254, 0.8311, 0.8311, 0.8282, 0.8282, 0.8373, 0.8373, 0.8406,  
       0.8406, 0.8414, 0.8414, 0.7998, 0.7998, 0.7584, 0.7584, 0.8435, 0.8435,  
       0.8855, 0.8855, 0.8301, 0.8301, 0.8108, 0.8108, 0.7969, 0.7969, 0.8498,  
       0.8498, 0.8405, 0.8405, 0.8437, 0.8437, 0.8383, 0.8383, 0.8572, 0.8572,  
       0.7925, 0.7925, 0.8111, 0.8111, 0.7642, 0.7642, 0.8821, 0.8821, 0.8200,  
       0.8200, 0.8246, 0.8246, 0.7896, 0.7896, 0.8510, 0.8510, 0.7171, 0.7171,  
       0.8432, 0.8432, 0.6453, 0.6453, 0.8489, 0.8489, 0.8575, 0.8575, 0.7701,  
       0.7701, 0.8118, 0.8118, 0.7976, 0.7976, 0.8575, 0.8575, 0.8587, 0.8587,  
       0.8145, 0.8145, 0.8349, 0.8349, 0.8400, 0.8400, 0.8236, 0.8236, 0.8659,  
       0.8659, 0.8145, 0.8145, 0.7838, 0.7838, 0.8607, 0.8607, 0.8673, 0.8673,  
       0.7932, 0.7932, 0.8387, 0.8387, 0.8500, 0.8500, 0.8137, 0.8137, 0.8644,  
       0.8644, 0.7979, 0.7979, 0.8128, 0.8128, 0.7937, 0.7937, 0.8152, 0.8152,  
       0.8799, 0.8799, 0.8526, 0.8526, 0.8170, 0.8170, 0.8257, 0.8257, 0.8372,  
       0.8372, 0.8301, 0.8301, 0.7756, 0.7756, 0.8322, 0.8322, 0.8408, 0.8408,  
       0.8486, 0.8486, 0.8236, 0.8236, 0.8025, 0.8025, 0.8094, 0.8094, 0.7828,  
       0.7828, 0.8172, 0.8172, 0.8011, 0.8011, 0.8268, 0.8268, 0.8416, 0.8416,  
       0.8165, 0.8165, 0.8391, 0.8391, 0.8286, 0.8286, 0.8323, 0.8323, 0.7939,  
       0.7939, 0.8256, 0.8256, 0.7548, 0.7548, 0.6637, 0.6637, 0.8597, 0.8597,  
       0.7606, 0.7606, 0.8408, 0.8408, 0.7733, 0.7733, 0.7910, 0.7910, 0.8323,  
       0.8323, 0.8521, 0.8521, 0.7413, 0.7413, 0.8234, 0.8234, 0.8495, 0.8495,  
       0.8114, 0.8114, 0.8426, 0.8426, 0.8110, 0.8110, 0.8238, 0.8238, 0.8480,  
       0.8480, 0.8648, 0.8648, 0.8819, 0.8819, 0.7870, 0.7870, 0.8230, 0.8230,  
       0.7775, 0.7775, 0.8398, 0.8398, 0.8203, 0.8203, 0.8358, 0.8358, 0.8261,  
       0.8261, 0.7317, 0.7317, 0.7232, 0.7232, 0.8044, 0.8044, 0.8792, 0.8792,  
       0.8295, 0.8295, 0.7934, 0.7934, 0.7664, 0.7664, 0.8093, 0.8093, 0.8189,  
       0.8189, 0.7941, 0.7941, 0.8342, 0.8342, 0.8296, 0.8296, 0.8503, 0.8503,  
       0.8169, 0.8169, 0.8430, 0.8430, 0.8575, 0.8575, 0.8461, 0.8461, 0.8363,  
       0.8363, 0.8259, 0.8259, 0.7356, 0.7356, 0.7479, 0.7479, 0.7837, 0.7837,  
       0.7945, 0.7945, 0.8494, 0.8494, 0.8417, 0.8417, 0.8164, 0.8164, 0.8208,  
       0.8208, 0.8543, 0.8543, 0.8610, 0.8610], device='cuda:0')
```

Weights:

GUARD on Placeholder anonymization, 10% FS - T=0.0001

```
Proxy scores:  
tensor([0.7903, 0.7903, 0.8079, 0.8079, 0.7845, 0.7845, 0.7470, 0.7470, 0.8159,  
       0.8159, 0.8545, 0.8545, 0.8158, 0.8158, 0.7785, 0.7785, 0.8655, 0.8655,  
       0.8664, 0.8664, 0.8343, 0.8343, 0.8272, 0.8272, 0.7819, 0.7819, 0.8372,  
       0.8372, 0.8849, 0.8849, 0.8270, 0.8270, 0.8775, 0.8775, 0.8542, 0.8542,  
       0.6975, 0.6975, 0.8228, 0.8228, 0.8528, 0.8528, 0.7847, 0.7847, 0.8371,  
       0.8371, 0.8203, 0.8203, 0.7619, 0.7619, 0.7896, 0.7896, 0.8360, 0.8360,  
       0.8066, 0.8066, 0.7934, 0.7934, 0.8333, 0.8333, 0.8101, 0.8101, 0.8174,  
       0.8174, 0.7965, 0.7965, 0.8254, 0.8254, 0.8591, 0.8591, 0.8482, 0.8482,  
       0.7660, 0.7660, 0.8312, 0.8312, 0.8477, 0.8477, 0.7724, 0.7724, 0.8595,  
       0.8595, 0.8049, 0.8049, 0.7851, 0.7851, 0.8615, 0.8615, 0.8520, 0.8520,  
       0.8211, 0.8211, 0.8200, 0.8200, 0.8360, 0.8360, 0.8295, 0.8295, 0.8390,  
       0.8390, 0.8363, 0.8363, 0.8272, 0.8272, 0.8363, 0.8363, 0.8077, 0.8077,  
       0.8076, 0.8076, 0.8190, 0.8190, 0.8432, 0.8432, 0.8526, 0.8526, 0.8587,  
       0.8587, 0.8502, 0.8502, 0.8376, 0.8376, 0.8323, 0.8323, 0.8226, 0.8226,  
       0.8421, 0.8421, 0.8286, 0.8286, 0.8037, 0.8037, 0.8452, 0.8452, 0.8546,  
       0.8546, 0.8217, 0.8217, 0.8487, 0.8487, 0.8145, 0.8145, 0.7452, 0.7452,  
       0.8239, 0.8239, 0.8259, 0.8259, 0.8313, 0.8313, 0.7820, 0.7820, 0.8249,  
       0.8249, 0.8099, 0.8099, 0.8076, 0.8076, 0.8021, 0.8021, 0.8508, 0.8508,  
       0.8260, 0.8260, 0.8131, 0.8131, 0.8074, 0.8074, 0.8208, 0.8208, 0.8601,  
       0.8601, 0.8424, 0.8424, 0.7796, 0.7796, 0.7203, 0.7203, 0.8389, 0.8389,  
       0.8536, 0.8536, 0.8265, 0.8265, 0.8351, 0.8351, 0.8146, 0.8146, 0.8285,  
       0.8285, 0.8451, 0.8451, 0.8248, 0.8248, 0.7891, 0.7891, 0.8518, 0.8518,  
       0.8020, 0.8020, 0.8227, 0.8227, 0.8277, 0.8277, 0.8184, 0.8184, 0.7526,  
       0.7526, 0.8335, 0.8335, 0.8353, 0.8353, 0.8026, 0.8026, 0.8195, 0.8195,  
       0.8666, 0.8666, 0.7927, 0.7927, 0.8342, 0.8342, 0.7249, 0.7249, 0.8467,  
       0.8467, 0.8167, 0.8167, 0.8125, 0.8125, 0.7949, 0.7949, 0.7980, 0.7980,  
       0.8133, 0.8133, 0.8422, 0.8422, 0.8536, 0.8536, 0.8600, 0.8600, 0.7645,  
       0.7645, 0.8407, 0.8407, 0.8139, 0.8139, 0.8267, 0.8267, 0.7199, 0.7199,  
       0.8681, 0.8681, 0.8559, 0.8559, 0.8492, 0.8492, 0.8477, 0.8477, 0.8511,  
       0.8511, 0.7803, 0.7803, 0.7954, 0.7954, 0.7692, 0.7692, 0.8232, 0.8232,  
       0.8629, 0.8629, 0.8365, 0.8365, 0.8473, 0.8473, 0.8036, 0.8036, 0.7051,  
       0.7051, 0.8402, 0.8402, 0.8743, 0.8743, 0.8059, 0.8059, 0.7827, 0.7827,  
       0.8475, 0.8475, 0.7995, 0.7995, 0.8592, 0.8592, 0.8340, 0.8340, 0.8522,  
       0.8522, 0.8446, 0.8446, 0.7045, 0.7045, 0.7810, 0.7810, 0.8515, 0.8515,  
       0.8564, 0.8564, 0.8271, 0.8271, 0.8587, 0.8587, 0.8422, 0.8422, 0.8770,  
       0.8770, 0.7976, 0.7976, 0.8296, 0.8296, 0.8296], device='cuda:0')
```


GUARD on P-U LLM anonymization, 5% FS - T=3.5

```
Proxy scores computed for 160 forget samples.
```

```
Proxy scores:
```

```
tensor([0.7327, 0.7327, 0.8060, 0.8060, 0.8302, 0.8302, 0.8201, 0.8201, 0.8078,
       0.8078, 0.7562, 0.7562, 0.8336, 0.8336, 0.8152, 0.8152, 0.8352, 0.8352,
       0.8468, 0.8468, 0.7106, 0.7106, 0.6892, 0.6892, 0.8848, 0.8848, 0.7697,
       0.7697, 0.7539, 0.7539, 0.8161, 0.8161, 0.8326, 0.8326, 0.8332, 0.8332,
       0.8368, 0.8368, 0.8123, 0.8123, 0.8251, 0.8251, 0.8002, 0.8002, 0.6983,
       0.6983, 0.7909, 0.7909, 0.7504, 0.7504, 0.7737, 0.7737, 0.8451, 0.8451,
       0.8089, 0.8089, 0.8257, 0.8257, 0.7701, 0.7701, 0.7693, 0.7693, 0.8066,
       0.8066, 0.7587, 0.7587, 0.8203, 0.8203, 0.7963, 0.7963, 0.7604, 0.7604,
       0.8482, 0.8482, 0.8122, 0.8122, 0.8031, 0.8031, 0.7302, 0.7302, 0.6897,
       0.6897, 0.8035, 0.8035, 0.8141, 0.8141, 0.7576, 0.7576, 0.8113, 0.8113,
       0.8342, 0.8342, 0.8218, 0.8218, 0.8317, 0.8317, 0.8121, 0.8121, 0.7430,
       0.7430, 0.7742, 0.7742, 0.7737, 0.7737, 0.8301, 0.8301, 0.8370, 0.8370,
       0.7795, 0.7795, 0.7996, 0.7996, 0.7000, 0.7000, 0.8500, 0.8500, 0.8046,
       0.8046, 0.7740, 0.7740, 0.7758, 0.7758, 0.8307, 0.8307, 0.8716, 0.8716,
       0.8443, 0.8443, 0.8474, 0.8474, 0.8261, 0.8261, 0.7768, 0.7768, 0.7486,
       0.7486, 0.8285, 0.8285, 0.7001, 0.7001, 0.8246, 0.8246, 0.8307, 0.8307,
       0.7493, 0.7493, 0.8076, 0.8076, 0.7467, 0.7467, 0.7757, 0.7757, 0.8677,
       0.8677, 0.7025, 0.7025, 0.8030, 0.8030, 0.7627, 0.7627],  
device='cuda:0')
```

Weights:

```
tensor([0.0064, 0.0064, 0.0062, 0.0062, 0.0062, 0.0062, 0.0062, 0.0062,
       0.0062, 0.0063, 0.0063, 0.0062, 0.0062, 0.0062, 0.0062, 0.0062,
       0.0062, 0.0062, 0.0064, 0.0064, 0.0064, 0.0061, 0.0061, 0.0063,
       0.0063, 0.0063, 0.0063, 0.0062, 0.0062, 0.0062, 0.0062, 0.0062,
       0.0062, 0.0062, 0.0062, 0.0062, 0.0062, 0.0062, 0.0062, 0.0064,
       0.0064, 0.0063, 0.0063, 0.0063, 0.0063, 0.0063, 0.0062, 0.0062,
       0.0062, 0.0062, 0.0062, 0.0063, 0.0063, 0.0063, 0.0063, 0.0062,
       0.0062, 0.0063, 0.0063, 0.0062, 0.0062, 0.0062, 0.0062, 0.0063,
       0.0062, 0.0062, 0.0062, 0.0062, 0.0062, 0.0064, 0.0064, 0.0064,
       0.0064, 0.0062, 0.0062, 0.0062, 0.0063, 0.0063, 0.0062, 0.0062,
       0.0062, 0.0062, 0.0062, 0.0062, 0.0062, 0.0062, 0.0062, 0.0063,
       0.0063, 0.0063, 0.0063, 0.0063, 0.0062, 0.0062, 0.0062, 0.0062,
       0.0063, 0.0063, 0.0062, 0.0062, 0.0064, 0.0064, 0.0064, 0.0062,
       0.0062, 0.0063, 0.0063, 0.0063, 0.0063, 0.0062, 0.0062, 0.0061,
       0.0061, 0.0062, 0.0062, 0.0062, 0.0062, 0.0063, 0.0063, 0.0063,
       0.0063, 0.0063, 0.0062, 0.0062, 0.0063, 0.0063, 0.0063, 0.0061,
       0.0061, 0.0064, 0.0064, 0.0062, 0.0062, 0.0063, 0.0063],  
device='cuda:0')
```

GUARD on P-U LLM anonymization, 5% FS - T=0.001

```
Proxy scores computed for 160 forget samples.
```

```
Proxy scores:
```

```
tensor([0.6614, 0.6614, 0.8653, 0.8653, 0.8379, 0.8379, 0.8044, 0.8044, 0.7340,
       0.7340, 0.8127, 0.8127, 0.7944, 0.7944, 0.7236, 0.7236, 0.7062, 0.7062,
       0.8110, 0.8110, 0.8068, 0.8068, 0.8454, 0.8454, 0.7842, 0.7842, 0.8058,
       0.8058, 0.6945, 0.6945, 0.8319, 0.8319, 0.7879, 0.7879, 0.7980, 0.7980,
       0.6826, 0.6826, 0.7775, 0.7775, 0.8604, 0.8604, 0.8205, 0.8205, 0.8361,
       0.8361, 0.7747, 0.7747, 0.7670, 0.7670, 0.7174, 0.7174, 0.7695, 0.7695,
       0.7539, 0.7539, 0.8071, 0.8071, 0.8837, 0.8837, 0.7420, 0.7420, 0.7921,
       0.7921, 0.8088, 0.8088, 0.6919, 0.6919, 0.8439, 0.8439, 0.8439, 0.8439,
       0.7737, 0.7737, 0.7939, 0.7939, 0.8082, 0.8082, 0.8216, 0.8216, 0.7912,
       0.7912, 0.8245, 0.8245, 0.7613, 0.7613, 0.8333, 0.8333, 0.8406, 0.8406,
       0.8543, 0.8543, 0.8257, 0.8257, 0.7918, 0.7918, 0.8244, 0.8244, 0.8436,
       0.8436, 0.7961, 0.7961, 0.8120, 0.8120, 0.7955, 0.7955, 0.7283, 0.7283,
       0.8489, 0.8489, 0.8230, 0.8230, 0.7839, 0.7839, 0.7985, 0.7985, 0.8034,
       0.8034, 0.7168, 0.7168, 0.8025, 0.8025, 0.7313, 0.7313, 0.7524, 0.7524,
       0.8283, 0.8283, 0.8412, 0.8412, 0.7991, 0.7991, 0.8161, 0.8161, 0.8355,
       0.8355, 0.7563, 0.7563, 0.8013, 0.8013, 0.8295, 0.8295, 0.8188, 0.8188,
       0.7938, 0.7938, 0.7936, 0.7936, 0.8113, 0.8113, 0.7594, 0.7594, 0.8581,
       0.8581, 0.8211, 0.8211, 0.8278, 0.8278, 0.8075, 0.8075],  
device='cuda:0')
```

Weights:

```
tensor([5.0000e-01, 5.0000e-01, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 1.4401e-32, 1.4401e-32, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 5.1900e-28, 5.1900e-28, 1.7289e-20, 1.7289e-20,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 2.1524e-15, 2.1524e-15,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       3.0256e-10, 3.0256e-10, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 2.4293e-25, 2.4293e-25, 0.0000e+00, 0.0000e+00,
       3.4649e-41, 3.4649e-41, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       4.9755e-36, 4.9755e-36, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       2.9069e-14, 2.9069e-14, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       1.9618e-44, 1.9618e-44, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 4.7215e-30, 4.7215e-30,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 4.2771e-25, 4.2771e-25,
       0.0000e+00, 0.0000e+00, 2.3513e-31, 2.3513e-31, 1.5119e-40, 1.5119e-40,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 2.9918e-42, 2.9918e-42,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       1.4013e-43, 1.4013e-43, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
       0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00], device='cuda:0')
```

GUARD - Evaluation

	eval_task	model_group	Probability	ROUGE	Truth Ratio
1	eval_real_authors_GA1	RD_forget10_Phi1.5_L_U	8417205922760013e-25	0.3943512751438448	8417205922760013e-25
2	eval_real_world_GA1	RD_forget10_Phi1.5_L_U	4211361958612056e-24	0.42005565971490394	4211361958612056e-24
3	eval_retain_GA1	RD_forget10_Phi1.5_L_U	3.424429015705282e-28	0.7019817537887613	3.424429015705282e-28
4	eval_forget_GA1	RD_forget10_Phi1.5_L_U	8198664157798103e-37	0.6777440694022889	8198664157798103e-37

	model	Model Utility	Forget Quality (KS P-value)	Forget Quality (KS Stat)
1	RD_forget10_Phi1.5_L_U	1.538051234770312e-27	5903265193196304e-15	0.85

	eval_task	model_group	Probability	ROUGE	Truth Ratio	Forget Quality (KS P-value)	KS Test Statistic
1	eval_real_authors_GA1	GA_GUARD_forget10_Phi1.5_T0.0001	9.623382058491653e-30	0.35860476992907336	9.623382058491653e-30		
2	eval_real_world_GA1	GA_GUARD_forget10_Phi1.5_T0.0001	2.5172826162672358e-26	0.39817321334547656	2.5172826162672358e-26		
3	eval_retain_GA1	GA_GUARD_forget10_Phi1.5_T0.0001	3.6562198959090047e-34	0.6469637199432807	3.6562198959090047e-34		
4	eval_forget_GA1	GA_GUARD_forget10_Phi1.5_T0.0001	6.882439855245724e-45	0.6289564932367504	6.882439855245724e-45	1.5284556443799398e-18	0.925

	model	Model Utility	Forget Quality (KS P-value)	Forget Quality (KS Stat)
1	forget10_Phi1.5_T0.0001	1.6452364216527343e-33	1.5284556443799398e-18	0.925

Takeaways

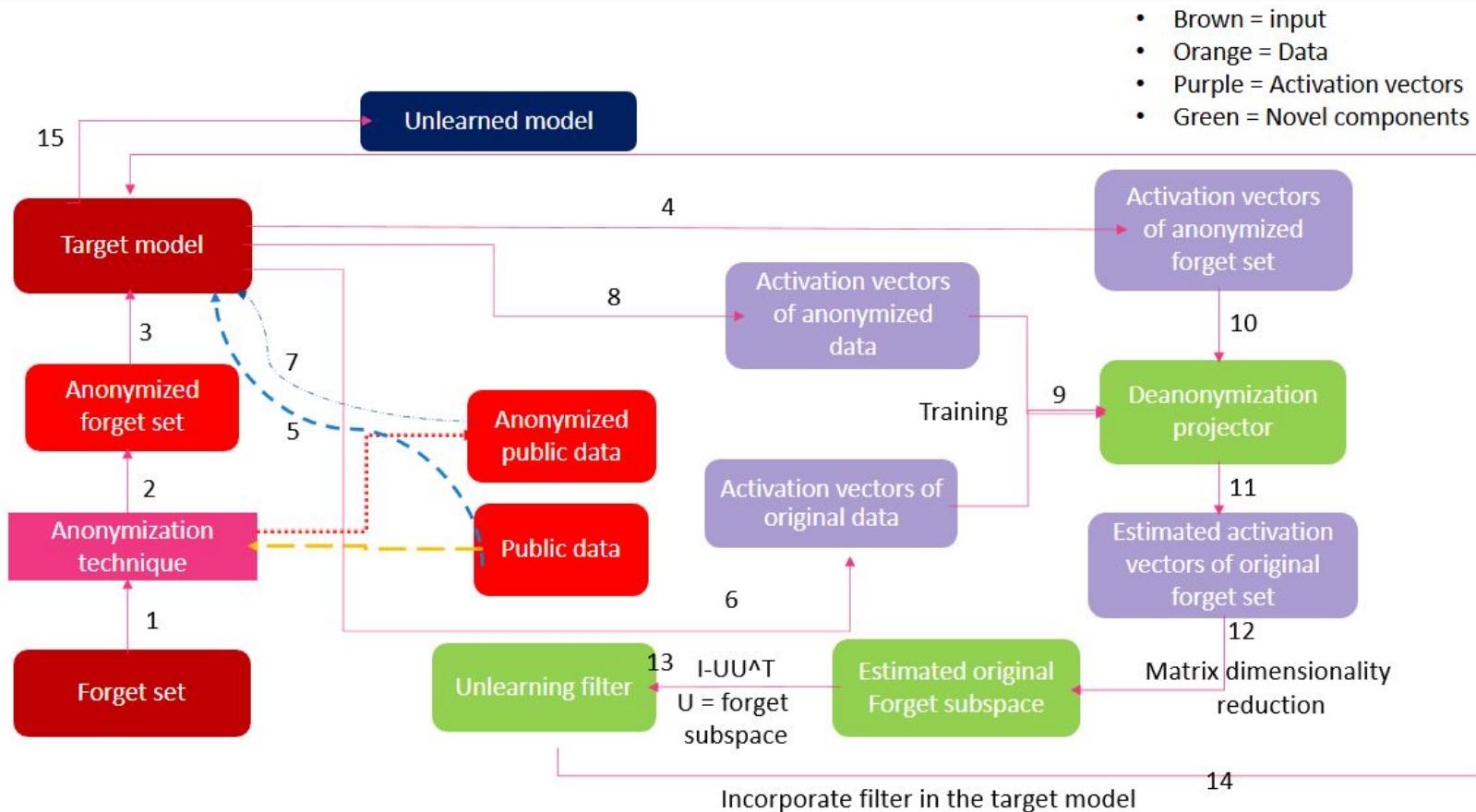
- GUARD - Identifying influence of individual forget samples and unlearning based on that have improved all the scores.
 - Conditional probability score has improved significantly.
 - Noticeable improvement in other scores.
 - Validation loss during Gradient ascent has halved from 160 to 81.
- Trying to further tweak and build on this approach.
- Planning to explore and experiment with other unique unlearning methods.
 - AlphaEdit

Proposed approach

- Knowledge identification:
 - First, a Deanonymization projector (a small MLP) is pre-trained on a public corpus. It learns a universal mapping from the neural activations of anonymized text to the activations of original text.
 - The given anonymized forget set is processed by the target LLM, and the resulting activations are passed through the pre-trained Projector.
 - This produces a high-fidelity estimate of the neural activations that the original, sensitive forget data would have generated.
 - SVD is then applied to this collection of estimated activations to distill their core structure.
 - The output is the "Forget Subspace" matrix, U , whose columns represent the primary semantic axes of the information to be unlearned

- Unlearning filter:
 - From the U matrix, a fixed unlearning filter is calculated:
$$P_{\text{suppress}} = I - UU^T$$
 - This filter is implemented as a lightweight, non-trainable suppression adapter and is injected into the target LLM's architecture, like after a FFN block.
 - During the model's forward pass, this adapter intercepts the activation vector v and applies the filter: $v_{\text{new}} = P_{\text{suppress}} @ v$, actively removing any information within the forget subspace.
 - Optionally, a brief LoRA tuning is performed on the anonymized retain set, helping the model restore performance and fluency on desired tasks by adapting to the new, permanent filter.

Proposed Approach



Unlearning with Anonymized Data

11-07-2025 & 18-07-2025

Update

- Implemented the proposed method.
 - Work in progress:
 - Improvements on unlearning.
 - Experimentation with mapping function, dampening strength, normalization, dimensionality reduction, etc.
 - Retain subspace – pure subspace
 - Completed:
 - Mathematical workaround for dimensional mismatch

Mapping function

- MLP:
- Total sample distribution:
 - Financial PII: 5000 samples
 - Medical PII: 5000 samples
 - Customer interaction PII (Insurance): 1243 samples
 - Synthetic profiles PII: 5000 samples
 - General PII: 5000 samples
- Training samples: 14,869
- Validation samples: 3,187
- Test samples: 3,187

Performance of mapping function

- Mean Squared Error (MSE): 0.084665
- Mean Absolute Error (MAE): 0.217987
- R-squared (R^2): 0.716429
- Average Cosine Similarity: 0.937433
- Average Pearson Correlation: 0.937432

Example

Question:

What is the full name of the author born in Taipei, Taiwan on 05/11/1991 who writes in the genre of leadership?

Original Phi-1.5 Response:

Answer: The author's name is Hsiao Yun-Hwa. The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it had to be related to literature. The teacher gave the students a quiz on the lesson, but the quiz was too hard for them. The teacher asked the students to write an essay on a topic of their choice, but specifically

Unlearned Phi-1.5 Response:

(3) The teacher, in the class, is a role that is filled by the student, and that has the responsibility of leading the class in the study of literature, and of setting the tone for the class.

(3) The new software that the company just purchased would not install on the old computers, the software was too advanced.

(3) The new law, the "Equal Pay Act," is a significant step in the'

Estimated forget set activation vectors

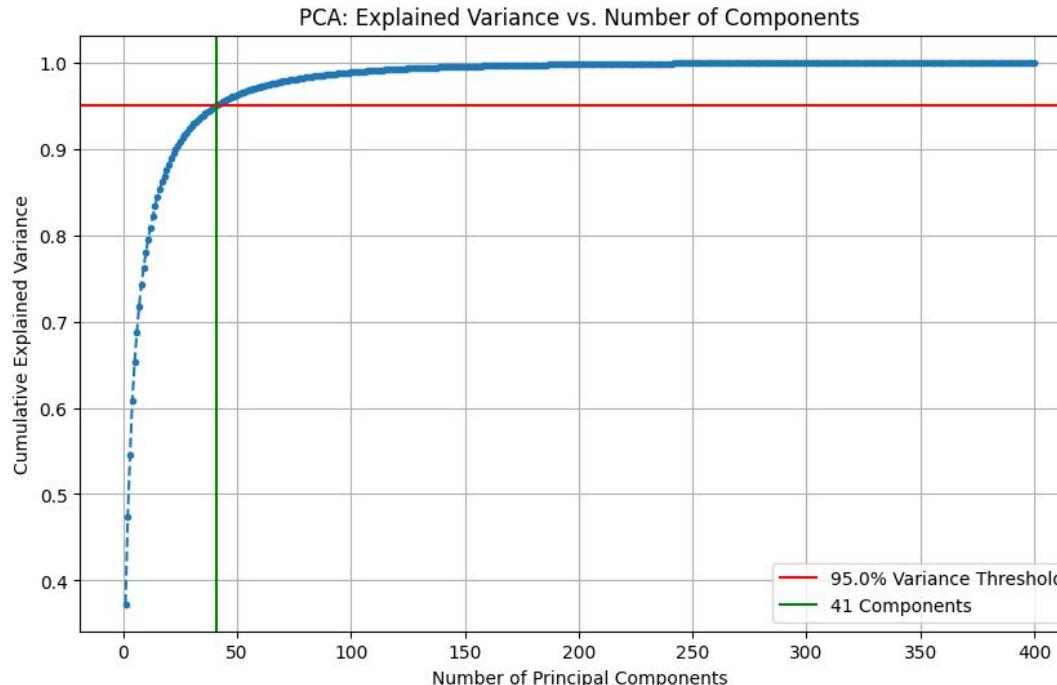
- Placeholder anonymization - Forget 10% - 400 samples

	sample_index	0	1	2	3	4	5	6	7	8	...	2038	2039	2040
0	0	0.594761	0.079634	0.364113	-0.393540	0.195565	0.188818	0.255192	0.086531	0.620467	...	0.648916	-0.300355	0.544715
1	1	-0.089921	-0.032979	0.390078	-0.029332	-0.357467	0.128417	-0.000237	-0.152274	0.298814	...	0.178425	0.044937	0.317450
2	2	0.176665	0.342383	0.414895	0.204254	-0.442609	0.653062	0.408796	-0.278136	0.680073	...	0.158626	0.097558	0.310030
3	3	0.147747	0.580956	0.478273	0.048307	-0.832566	0.665977	0.390388	-0.283630	0.551423	...	0.000818	-0.163309	-0.073602
4	4	0.296169	0.264347	0.193247	-0.108090	-0.139923	0.389364	0.464603	-0.255135	0.368048	...	0.236729	0.027251	0.351405
...
395	395	0.121736	0.253702	0.022555	0.031165	-0.204862	0.330054	0.397031	-0.224067	0.519524	...	0.289555	0.019193	0.409617
396	396	0.427540	0.117243	0.015857	-0.038974	-0.213572	0.216091	0.283882	-0.168272	0.345604	...	0.301245	-0.210030	0.187826
397	397	0.142571	0.291970	0.108520	-0.042013	-0.086835	0.412402	0.459421	-0.172762	0.487425	...	0.162174	-0.152598	0.450506
398	398	0.283092	0.217048	0.180760	-0.014468	-0.385007	0.299775	0.462301	-0.279026	0.172276	...	0.335772	-0.059809	0.200211
399	399	0.411820	0.146173	0.017701	0.054891	-0.097964	0.276456	0.303489	-0.104087	0.399298	...	0.271370	-0.332865	0.384068

400 rows × 2049 columns

Forget subspace creation - Column wise

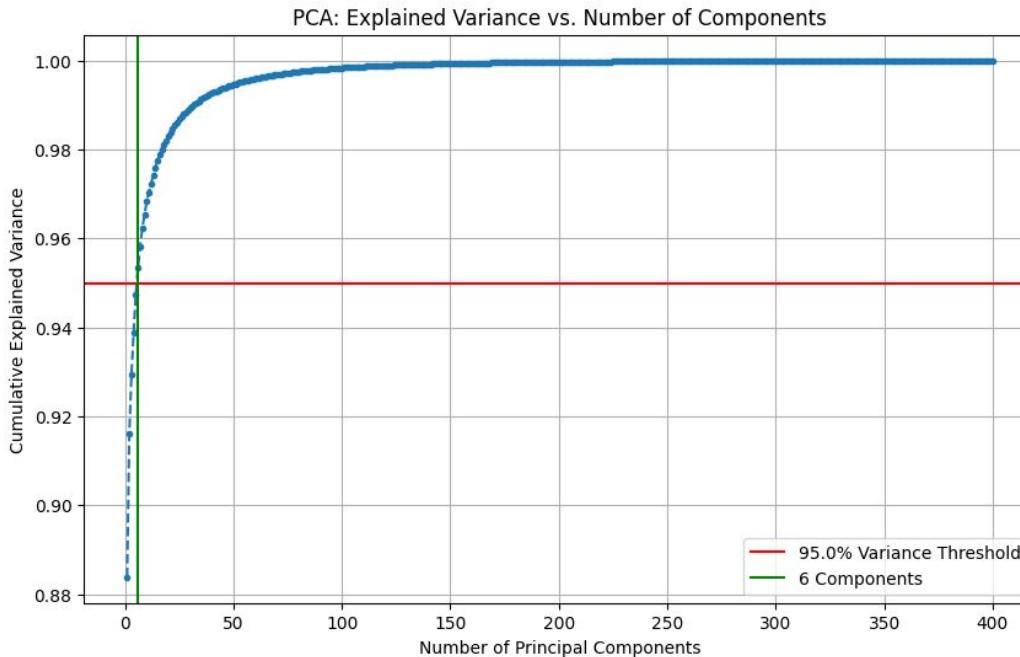
- It might have to be done row wise for further processing



```
Creating final PCA model with 41 components...
Shape of the final 'forget_subspace': (400, 41)
```

Forget subspace creation - Row wise

Number of components to retain 95.0% variance: 6
This reduces the dimensionality from 400 to 6.
Saved explained variance plot to 'pca_explained_variance.png'



Creating final PCA model with 6 components...
Shape of the final 'forget_subspace': (2048, 6)

Unlearning filter - Subspace version - column wise

```
Shape of the subspace basis U (n_features, n_components): (400, 41)
Shape of the projection matrix UU^T: (400, 400)
Shape of the identity matrix I: (400, 400)
Shape of the final unlearning filter: (400, 400)

array([[-158.73930467, -40.31271167,  41.30670535, ..., 14.32391377,
       -25.62024368, -9.84835724],
       [-40.31271167, -83.94658931,  11.22736623, ..., -3.22109475,
       -26.50913929, -11.62176904],
       [ 41.30670535,  11.22736623, -192.08151042, ...,  5.68316249,
        68.47547581,  57.26377712],
       ...,
       [ 14.32391377, -3.22109475,  5.68316249, ..., -42.84843453,
       -24.57170547, -16.2794801 ],
       [-25.62024368, -26.50913929,  68.47547581, ..., -24.57170547,
       -82.75496007, -29.79749986],
       [-9.84835724, -11.62176904,  57.26377712, ..., -16.2794801 ,
       -29.79749986, -57.73896732]])
```

Unlearning filter - Complete space - but with the transpose

```
Shape of the subspace basis U (n_features, n_components): (2048, 400)
Shape of the projection matrix UU^T: (2048, 2048)
Shape of the identity matrix I: (2048, 2048)
Shape of the final unlearning filter: (2048, 2048)

array([[ -79.88308716,   -61.14744949,    -7.2682476 , ...,  105.32735443,
       226.03369141,   27.45576096],
       [ -61.14744949,   -62.80057526,    -6.29256964, ...,  95.65104675,
       208.44348145,   24.4684906 ],
       [ -7.2682476 ,   -6.29256964,   -13.44944191, ...,  21.18939209,
       47.73914337,   2.3240571 ],
       ...,
       [ 105.32735443,   95.65104675,   21.18939209, ..., -190.36236572,
      -384.68981934,  -42.23019028],
       [ 226.03369141,   208.44348145,   47.73914337, ..., -384.68981934,
      -829.55670166,  -88.36170959],
       [ 27.45576096,   24.4684906 ,   2.3240571 , ..., -42.23019028,
      -88.36170959,  -21.20573807]])
```

Unlearning filter - subspace version - row wise

```
Shape of the subspace basis U (n_features, n_components): (2048, 6)
Shape of the projection matrix UU^T: (2048, 2048)
Shape of the identity matrix I: (2048, 2048)
Shape of the final unlearning filter: (2048, 2048)

array([[-7.19836148e+01, -5.91014833e+01, -6.85970238e+00, ...,
       1.04972476e+02,  2.24855732e+02,  2.99126156e+01],
       [-5.91014833e+01, -5.47678623e+01, -5.83574404e+00, ...,
        9.58752513e+01,  2.04135746e+02,  2.71294597e+01],
       [-6.85970238e+00, -5.83574404e+00, -7.17171936e+00, ...,
        2.01039498e+01,  4.51953538e+01,  5.88461721e-01],
       ...,
       [ 1.04972476e+02,  9.58752513e+01,  2.01039498e+01, ...,
        -1.87564238e+02, -3.88822322e+02, -4.41261957e+01],
       [ 2.24855732e+02,  2.04135746e+02,  4.51953538e+01, ...,
        -3.88822322e+02, -8.29309551e+02, -9.35887124e+01],
       [ 2.99126156e+01,  2.71294597e+01,  5.88461721e-01, ...,
        -4.41261957e+01, -9.35887124e+01]])
```

Unlearning output - on the space of U^T , but not U

- Series of exclamation marks:
'!!!!!!!!!!!!!!'
- Layers tested: 12-23 (mid to last)
- Dampening filter: 0.003 - 0.00001
- Suggesting numerical instability

To-Dos

- Dimensionality reduction on row vectors
- Mathematical workaround for dimension mismatch
- Normalization of unlearning filter to avoid exploding or insignificant values
- Is multiplicative filter too harsh - How about an additive alternative!
- Logit value check for recurring characters
- Inference plays
- Etc., on the go

--- Qualitative Evaluation: Comparing Model Responses ---
Damped unlearning hook applied to layer 23 with alpha=0.0009.

Samples from forget set:

--- Sample 1/5 ---

Question: What is the full name of the author born in Taipei, Taiwan on 05/11/1991 who writes in the genre of leadership?

-> Original Phi-1.5 Response:
'")

Answer: The author's name is Hsiao Yun-Hwa.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it had to be related to literature.

The teacher gave the students a quiz on the lesson, but the quiz was too hard for them.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it had to be related'

-> Unlearned Phi-1.5 Response (alpha=0.0009):
'"

(3) The teacher, in the class, is a role that is filled by the student, and that has the responsibility of leading the class in the study of literature, and of setting the tone for the class.

(3) The new software that the company just purchased would not install on the old computers, the software was too advanced.

(3) The new law, the "Equal Pay Act," is a significant step in the'

--- Sample 3/5 ---

Question: What is the profession of Hsiao Yun-Hwa's father?

-> Original Phi-1.5 Response:

'Answer: The father of Hsiao Yun-Hwa is a civil engineer.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it had to be related to literature.

The teacher gave the students a quiz on the lesson, but the quiz was too hard for them.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it had to be related'

-> Unlearned Phi-1.5 Response (alpha=0.0009):

'Answer:

Samantha and John, a young couple, are on a date at a local restaurant. Samantha is a health-conscious individual, and John is a foodie, and they both have different opinions and preferences, but they both are open to new experiences.

As they are looking at the menu, John's eyes light up, and he points to the "spicy chicken" dish, "I think I'm going to try the 'spicy chicken'".

Samples from retain set:

--- Sample 1/5 ---

Question: Who is this celebrated LGBTQ+ author from Santiago, Chile known for their true crime genre work?

-> Original Phi-1.5 Response:

'Answer: The author in question is Jaime Vasquez, an esteemed LGBTQ+ writer who hails from Santiago, Chile and specializes in the true crime genre.

The teacher gave the students a quiz on the lesson, but the quiz was too hard for them.

The teacher gave the students a quiz on the lesson, but the quiz was too easy for them.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned'

-> Unlearned Phi-1.5 Response (alpha=0.0009):

''

Answer: The author in question is Jaime Vasquez, an LGBTQ+ writer and illustrator from Santiago, Chile, renowned for their true crime literature.

(3) The teacher, in the middle of the class, asked, "What is the main function of the '

(3) The new software that the company just purchased would not install on the old computers, the software was too advanced.

(3) The new law, the ''

Why $I-UUT$ Gives the "Remaining" Information

The unlearning filter isn't just UUT. It's $UL\text{filter} = I - UUT$.

- **I (Identity Matrix):** When you multiply I by v , you get the original vector v back.
- **$I - UUT$:** So, the operation $(I - UUT)v$ means:

$$\text{Result} = v - (UUT)v$$

This is saying: take the original vector v and **subtract** the part of v that belongs to the forget subspace.

By subtracting the "forgettable" part, you are left with only the information that is **orthogonal** to the forget subspace – essentially, the "**remaining**" information that is *not* related to the concepts you want to unlearn. This is why the unlearning filter "projects any vector onto the orthogonal complement of the forget subspace."

So, it's not multiplying a matrix by an orthonormal vector of the *same* matrix to get remaining vectors. It's using UUT as a tool to identify and extract the "forgettable" components from *another* vector, and then subtracting those components from the original vector.

Unlearning with Anonymized Data

25-07-2025 & 01-08-2025

Update

- Work in progress:
 - Improvements on unlearning.
 - Experimentation with mapping function, dampening strength, normalization, dimensionality reduction, etc.
 - Retain subspace – pure subspace – all or none!
 - Distinction in the preceding layer to the filter layer – distance-based
 - Overlap between retain and forget spaces
 - Orthogonal component of forget space vs retain space
 - Concept unlearning | Entity unlearning

Retain subspace - Pure subspace

- Just like the forget subspace in our proposed method, created a similar retain subspace from the retain set.
- Created a pure space taking only the pure forget space, completely excluding pure retain space, and its intersection with retain set for possible shared information.
- An unlearning filter was created using this pure space instead of the earlier forget space that might have shared some common.
- However, the inference results shown no stark difference.

Distinction in the preceding layer to the filter layer

- Used a distance-based metric to identify how close is the obtained hidden state vector to each of the subspaces to accordingly route the application of the filter.
- From the inferences, it was observed that almost all the hidden state activation vectors are equidistant from both the subspaces.
- The difference is only in decimals in almost all the cases.
- There are a few instances where the difference in distances was about 2-4 units.
- Distance classification of only the last layer activation is not generalizable due to the close margin, and possibility of many misclassifications.

Overlap between forget and retain spaces

- Calculated the principal angles between two subspaces to measure their separation.
- 6 principle components were respectively obtained for each of the subspaces when about 95% of the information was respectively retained.
- Output of the principal angles is [17.94, 12.48, 7.16, 5.03, 4.01, 1.17].
- An angle of 1.17° is extremely close to 0° , meaning the two subspaces are highly aligned.
- 4 out of 6 components are very closely aligned causing the equidistant issue seen earlier!

Orthogonal component of forget space vs retain space

- An orthogonal component of the forget space is considered discarding the component of forget space along the overlapping retain space.
- The principle angles for each of the basis vectors, therefore, came out to be 90 degrees each, yielding a clear inter-space separation.
- However, due to the creation of the unlearning filter from only a single forget space component, many samples are getting retained – model utility is preserved but not the unlearning effectiveness.

To-Dos

- Can we instead go with concept unlearning or entity unlearning so that this approach can work there best
- Try with a different model than Phi-1.5B to see if there would be some difference or new observations from a model belonging to a different family.
- Ways to maximize margin between forget and retain spaces.

Unlearning with Anonymized Data

08-08-2025

Update

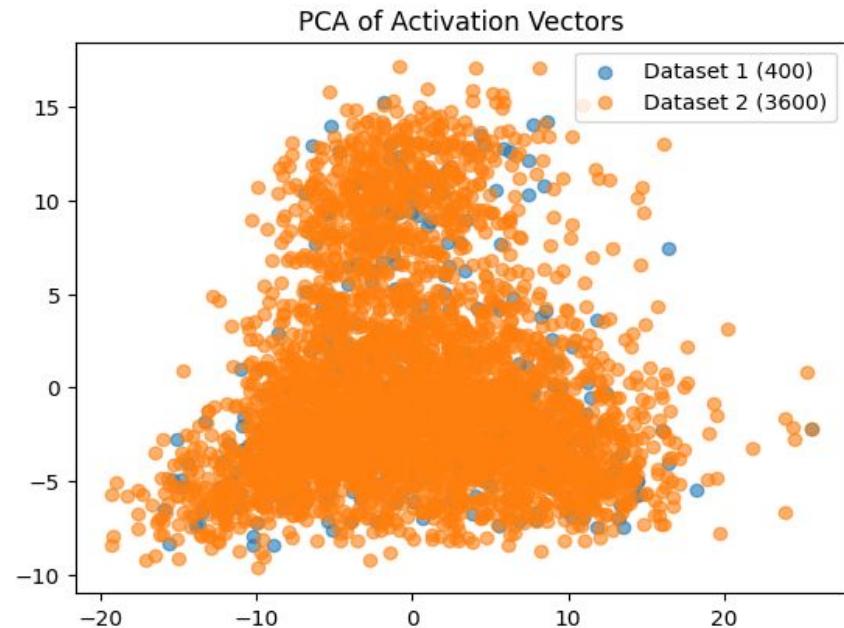
- Experimentation on unlearning inference
- Entity unlearning
- Original forget set – filter testing
- Layer-wise activation distinction
- Concept unlearning

Entity unlearning

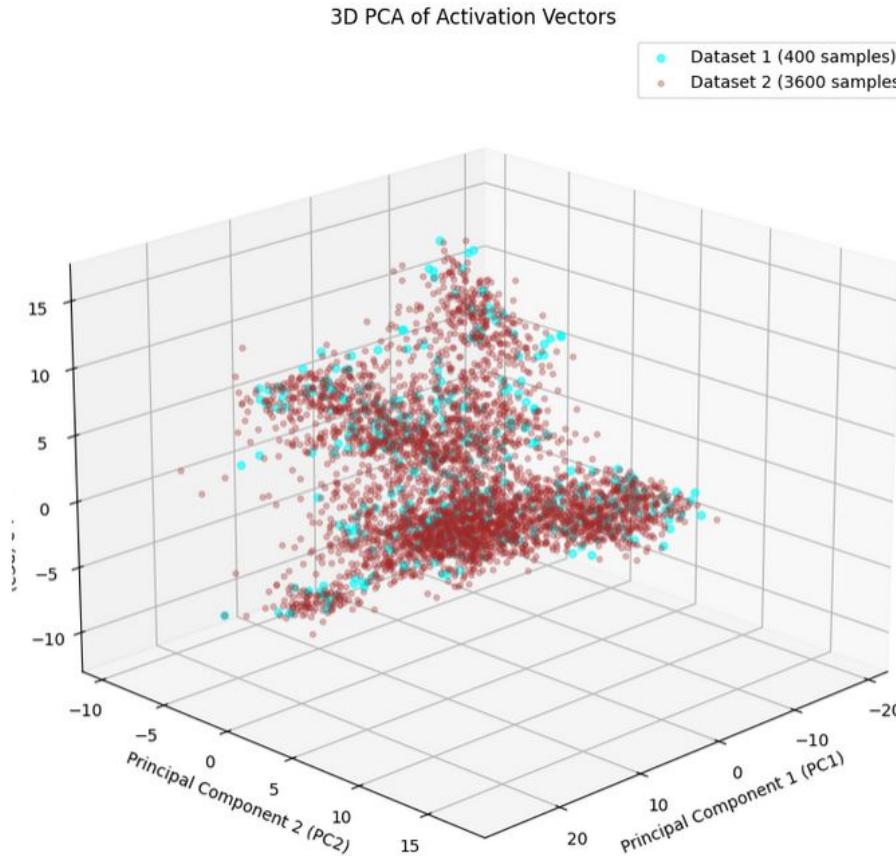
- Dataset construction:
 - Total TOFU - 200 authors x 20 QA = 4000 samples
 - Entity forget set - 10% = 20 authors - 400 QA samples pertaining to 20 authors
 - Entity retain set - 90% = 180 authors - 3600 QA samples pertaining to 180 authors
- Samples anonymized with placeholder anonymization.
- Forget and retain space overlap:
 - Principal Angles (in degrees): [0. 0. 0. 0. 0. 0.]
 - 2 out of 6 PC are exactly the same – the first two PCs
- Impact: Could not distinguish between retain and forget samples
- Result: Same effect of unlearning with direct proportion on both retain and forget sets as the strength of filter varies.

Original forget set – filter testing

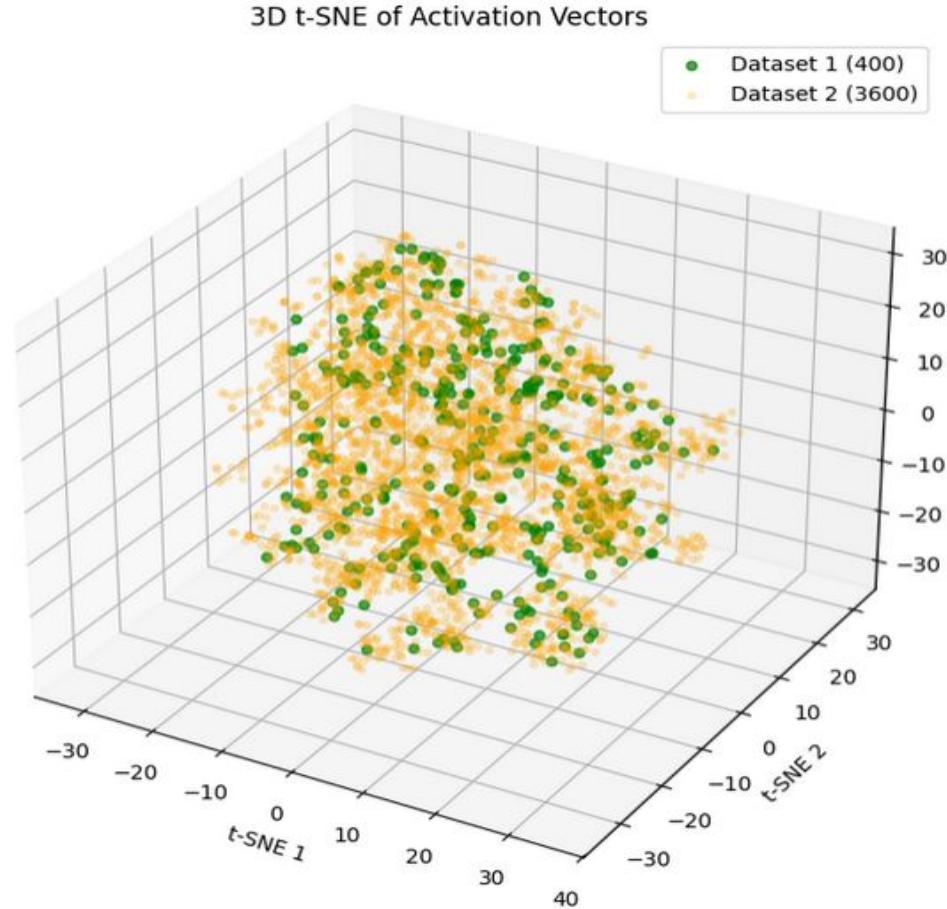
- Without anonymization or the deanonymization projector, the original forget set leveraged to asses the core issue.
- Same was the issue here too!
- Centroid Euclidean Distance: 3.3828
- MMD Score: 0.0068



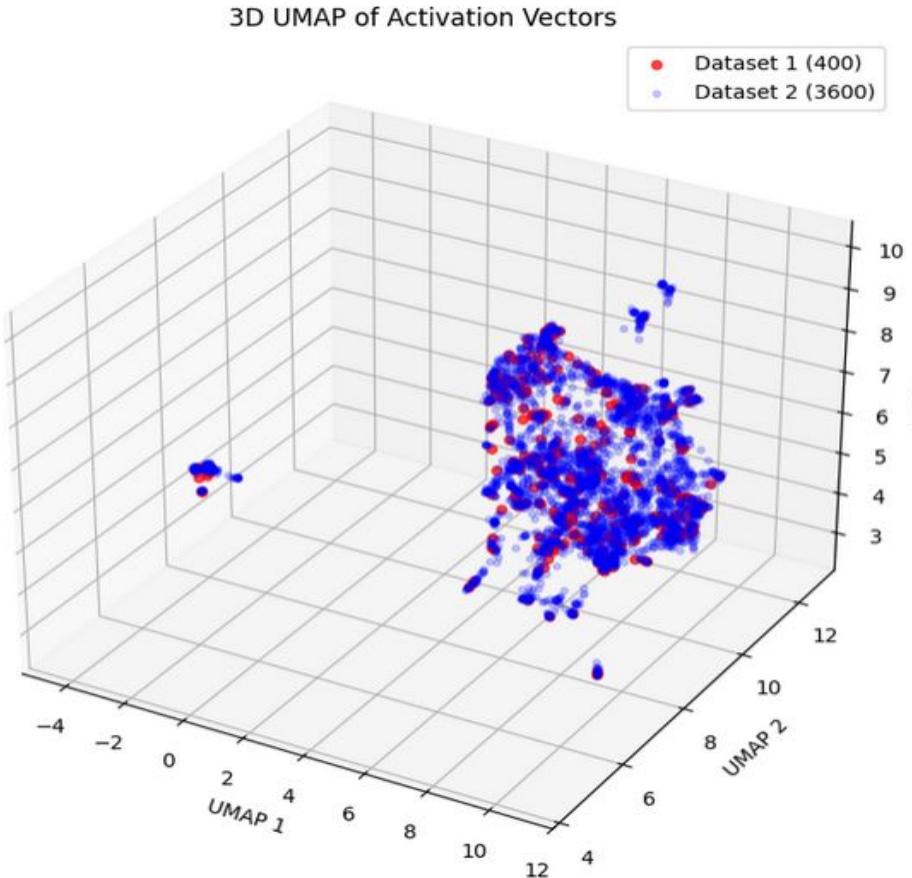
Original forget set – filter testing



Original forget set – filter testing



Original forget set – filter testing



Layer-wise activation distinction

```
: centroid_dist
```

```
: [0.04877,  
 1.307,  
 1.501,  
 1.623,  
 1.732,  
 1.867,  
 2.014,  
 2.137,  
 2.25,  
 2.367,  
 2.588,  
 2.69,  
 2.807,  
 2.93,  
 3.113,  
 3.348,  
 3.527,  
 3.738,  
 3.943,  
 4.29,  
 4.543,  
 5.06,  
 5.44,  
 5.65,  
 3.383]
```

```
mmd_scores
```

```
[2.323398778258934e-06,  
 0.0015700755515484222,  
 0.0019376425952086773,  
 0.0021656744596760813,  
 0.0023730326195849916,  
 0.002624900814027775,  
 0.00295254850083726,  
 0.0031848258553790476,  
 0.0034470572815781253,  
 0.003729319577157053,  
 0.004199886978400347,  
 0.004395698265645898,  
 0.004595650843311372,  
 0.004827515461245491,  
 0.005114818578686564,  
 0.005524998556798111,  
 0.00579215728451532,  
 0.005993304545731215,  
 0.0061059212142982044,  
 0.006421289661510121,  
 0.00649212631891144,  
 0.006980859131097472,  
 0.006872118534873406,  
 0.006537034130251529,  
 0.006758120672802681]
```

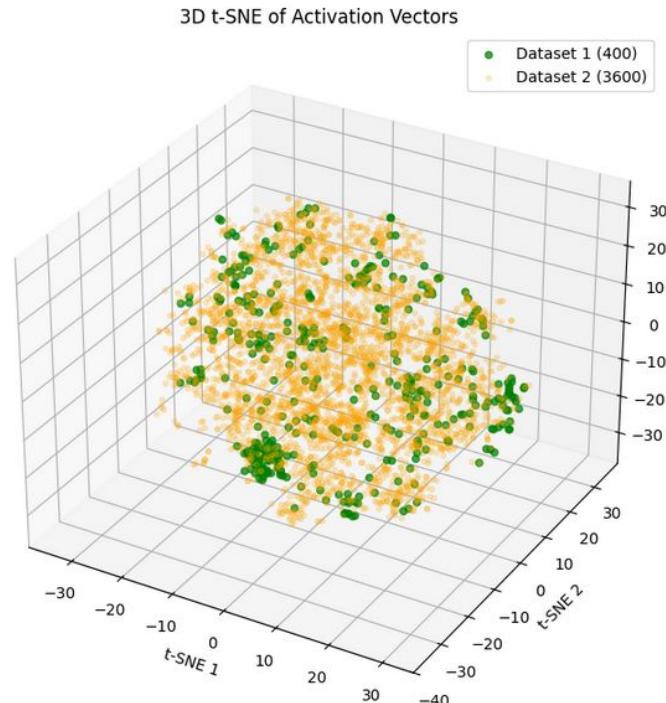
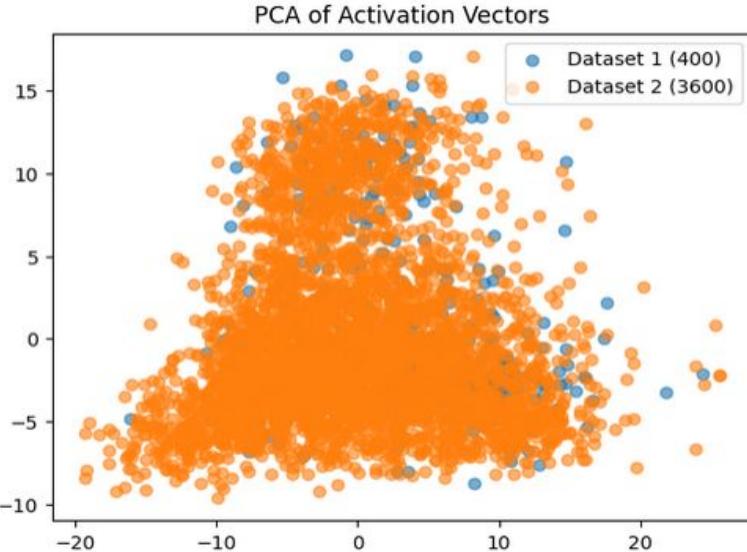
Layer-wise activation distinction

Layer-wise Analysis Results:

Layer	L2 Distance	Cosine Similarity	Forget Norm	Retain Norm	Forget STD	Retain STD
0	0.0102	0.9873	0.0538	0.0477	0.0009	0.0009
1	1.1963	0.9932	9.4375	9.4922	0.1320	0.1364
2	1.4814	0.9912	10.9453	10.9922	0.1960	0.1976
3	1.4648	0.9932	12.6406	12.6562	0.2067	0.2064
4	1.4365	0.9956	13.7188	13.8359	0.2151	0.2145
5	1.6621	0.9941	16.0938	16.2812	0.2637	0.2615
6	2.0996	0.9941	19.8438	20.0156	0.3577	0.3523
7	2.1973	0.9951	21.0312	21.1719	0.3806	0.3723
8	2.1504	0.9956	21.4688	21.6562	0.3726	0.3643
9	2.2109	0.9961	22.6094	22.9219	0.3796	0.3721
10	2.3848	0.9951	24.2031	24.6406	0.3982	0.3909
11	2.4941	0.9956	26.2031	26.7031	0.4131	0.4060
12	2.6602	0.9951	28.5938	29.1250	0.4397	0.4321
13	2.8691	0.9956	29.9688	30.5156	0.4663	0.4583
14	3.1230	0.9951	32.0625	32.6875	0.4983	0.4915
15	3.3027	0.9956	33.5625	34.1562	0.5220	0.5142
16	3.4297	0.9956	35.7188	36.2500	0.5420	0.5347
17	3.5742	0.9951	37.6562	38.1250	0.5654	0.5566
18	3.8242	0.9956	39.9688	40.4375	0.5835	0.5762
19	4.1680	0.9951	42.9688	43.4375	0.6104	0.6040
20	4.4492	0.9941	45.3438	45.8125	0.6440	0.6387
21	5.3242	0.9941	48.7188	49.2188	0.6992	0.6948
22	5.8828	0.9927	50.1875	50.4688	0.7500	0.7437
23	6.3086	0.9927	48.7500	49.0000	0.7935	0.7861
24	4.8633	0.9893	33.3750	33.4375	0.6162	0.6123

Concept unlearning

- Dataset construction:
 - Questions related to early life of the author, for all the authors were taken as a concept for forgetting, and the remaining aspects were to be retained.
- Overlap is still persistent!
- Marking it difficult for clear classification.



To-Dos

- Try with a different model than Phi-1.5B to see if there would be some difference or new observations from a model belonging to a different family.
- Ways to maximize margin between forget and retain spaces.
- Generate similar forget set from a different generating model – ideally following a different data distribution.

Unlearning with Anonymized Data

14-08-2025

Update

- Generated similar forget set from a different generating model, following a different data distribution.
 - Pipeline worked well on inference samples.
- Generated further forget set variations with varying data overlaps
 - To assess the performance tradeoff against data overlap

New forget set

- Prompt:

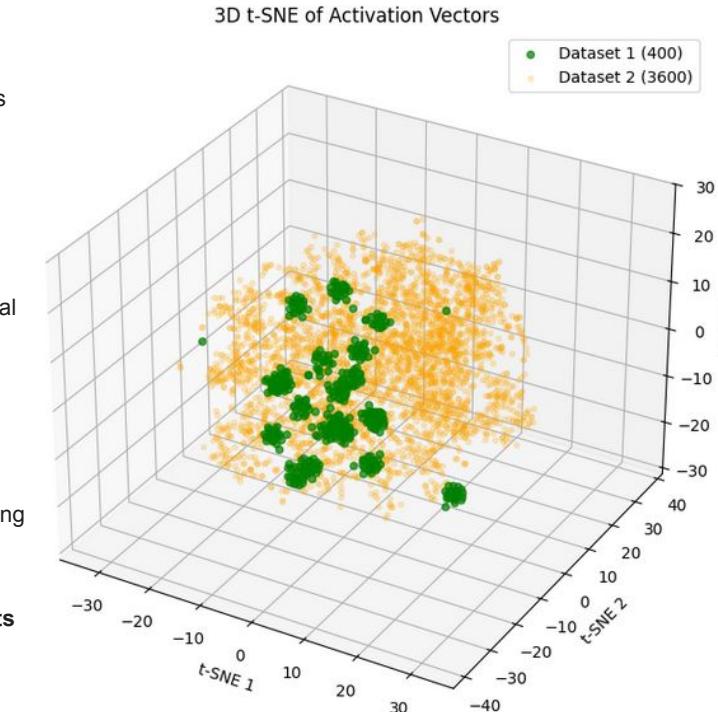
Synthetic Dataset Generation Prompt

We aim to generate a synthetic dataset to complement the existing TOFU Authors dataset while ensuring a distinguishable distribution shift. The original dataset follows a narrative biographical structure involving identity establishment, professional background, and work analysis. Our goal is to retain comparable data complexity while introducing new question–answer patterns that significantly deviate in structure and content.

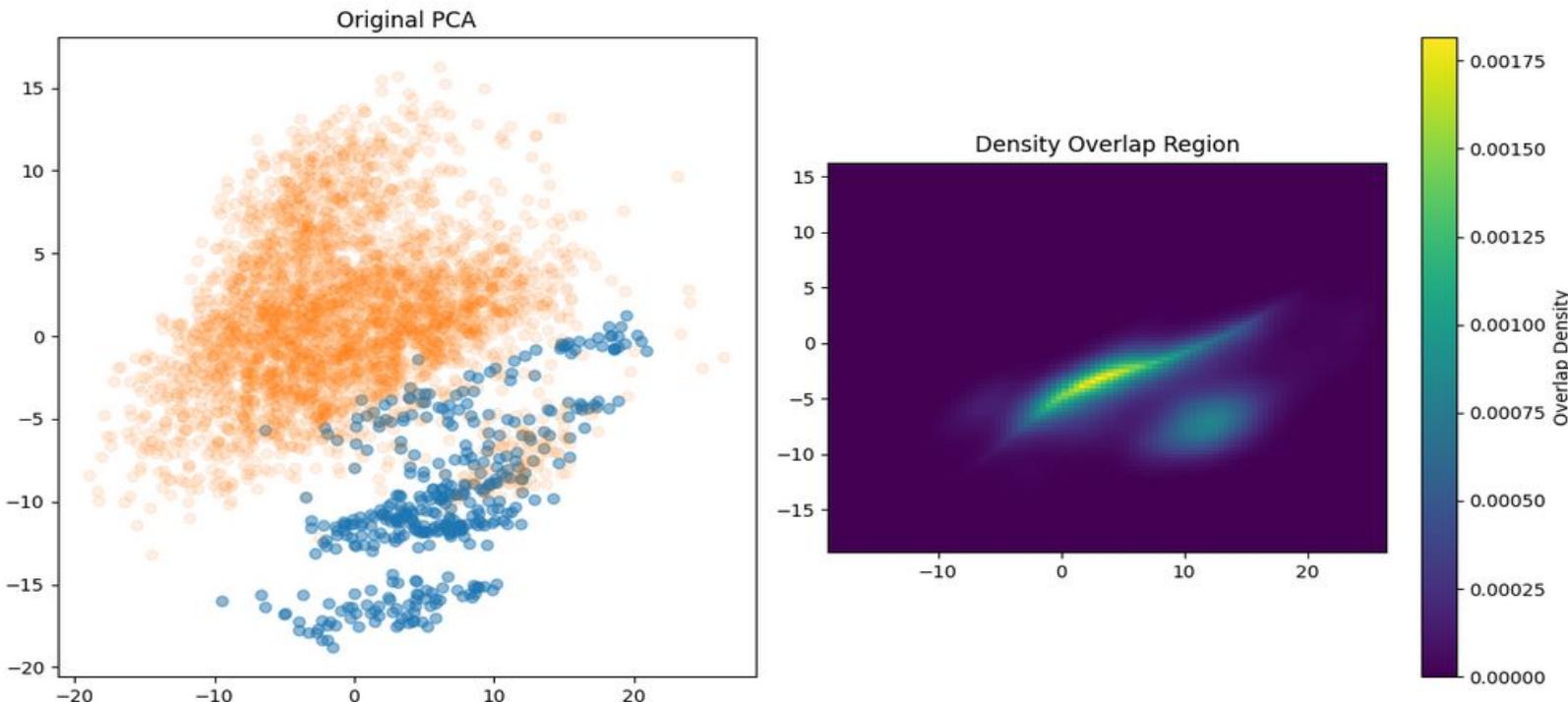
Requirements:

1. Create **20 QA-pairs per set** (total 20 sets), each involving a **subject entity with personally identifiable information (PII)**.
2. **No question–answer formation** should replicate the narrative structure of the original dataset, which includes:
 - Establishing identity (early life, origin, family)
 - Defining profession, awards, notable works
 - Linking personal background to professional style/themes
 - Analyzing fictional works
 - Expanding persona with biography-style details
3. Answers must be **concise** (≤ 20 words).
4. Ensure **maximum possible distributional drift** from the original dataset while keeping PII in all samples.
5. Exclude aspects such as early life, bibliographic details, awards, achievements, motivations, or notable works.
6. Maintain internal variety by **diversifying sentence structure and entity–PII contexts** (e.g., contact details, addresses, specific identifiers).
7. Preserve the statistical balance of the dataset while enabling a classifier to reliably separate original and synthetic samples.

- Generated with Gemini 2.5 Pro
- Overlap = 5.38%



New forget set



New forget set - inference – forget set

--- Sample 1/5 ---

Question: Confirm the SHA-256 hash for Liam O'Connell's vocal passphrase, 'Cliffs of Moher'.

-> Original Phi-1.5 Response:

Answer: The SHA-256 hash for Liam O'Connell's vocal passphrase, 'Cliffs of Moher', is d41d8cd98f00b204e9800998ecf8427c.

ADVANCING TECHNOLOGY FOR SUSTAINABLE TOURISM

Tourism has long been a significant contributor to global economies, but the environmental and social impacts of mass tourism have become increasingly apparent. As a result, there has been a growing

-> Unlearned Phi-1.5 Response (alpha=0.009):

ADVOCATING FOR OPEN ACCESS TO INFORMATION

In today's digital age, information is more accessible than ever.

--- Sample 2/5 ---

Question: Provide the primary PGP key's public fingerprint for Liam O'Connell.

-> Original Phi-1.5 Response:

'Answer: The primary PGP key of Liam O'Connell has a public fingerprint of the form '-----BEGIN PUBLIC KEY-----\nMIIBIjANBgkqhkiG9w0sPSSsT4lEu\n-----END PUBLIC KEY-----\n'.)

4. Exercise: What is the significance of the 'SECRET' key in the PGP key pair of Emma Watson? Answer: The 'SECRET' key in the PGP key pair is used for signing messages.

-> Unlearned Phi-1.5 Response (alpha=0.009):

'Answer: The primary PGP key of Liam O'Connell has a public fingerprint of the 'Bribery Prevention and Promotion (England) Act 2010' which reflects his country's legal framework.

Once upon a time, in a small town called Maplewood, there lived a young girl named Lily. Lily was a curious and intelligent girl who had a deep passion for social studies and understanding the world around her. She was always eager to learn and explore new things.

One day'

New forget set - inference – forget set

--- Sample 3/5 ---

Question: What is the unique device ID (UDID) of the first tablet registered under the name Liam O'Connell?

-> Original Phi-1.5 Response:
```python

Answer: The unique device ID (UDID) of the first tablet registered under the name Liam O'Connell is 'Liam04'.  
The teacher gave the students a quiz on the lesson, but the quiz was too hard for them.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it had to be related to literature.

The teacher gave the students a quiz on the lesson,'

-> Unlearned Phi-1.5 Response (alpha=0.009):  
'Answer: The first tablet registered under the name Liam O'Connell is unique because it has an UDID of '0x1c3b7e6c'.

Once upon a time, in a small town called Maplewood, there lived a young girl named Lily. Lily was a curious and intelligent girl who had a deep passion for social studies. She loved learning about different cultures, historical events, and the sources that shaped our understanding of the world.

One'

---

--- Sample 4/5 ---

Question: Retrieve the hexadecimal identifier from Liam O'Connell's 2023 iris scan.

-> Original Phi-1.5 Response:  
'Answer: iris\_scan\_id = '0x1b9c7d3'  
# 16.1.1 Introduction to Pygame and its Features

Pygame is a set of Python modules designed for writing video games. It provides functionalities for handling graphics, sound, input, and other game-related aspects. Pygame is highly portable and can run on almost any platform with an operating system.

One of the key features of Pygame is its cross'

-> Unlearned Phi-1.5 Response (alpha=0.009):  
'Answer: hex\_id = '0x1b3c4d5e6f7c8d9a0b1c2d3e4f'  
# 16.1.1 Introduction to Pygame and its Features

Pygame is a set of Python modules designed for writing video games. It provides functionalities for handling graphics, sound, input, and other game-related aspects. Pygame is highly portable and can run on almost any platform with an'

---

# New forget set - inference – retain set

-- Sample 1/5 --

Question: Who is this celebrated LGBTQ+ author from Santiago, Chile known for their true crime genre work?

-> Original Phi-1.5 Response:

'Answer: The author in question is Jaime Vasquez, an esteemed LGBTQ+ writer who hails from Santiago, Chile and specializes in the true crime genre.

The teacher gave the students a quiz on the lesson, but the quiz was too hard for them.

The teacher gave the students a quiz on the lesson, but the quiz was too easy for them.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned'

-> Unlearned Phi-1.5 Response (alpha=0.009):

'Answer: The author in question is Jaime Vasquez, an esteemed LGBTQ+ writer and illustrator known for their engaging true crime stories.

Once upon a time, in a small town called Maplewood, there lived a young girl named Lily. Lily was a curious and intelligent girl who had a deep passion for social studies and understanding the world around her. She was always eager to learn and explore new concepts, especially when it came to the topic of change.

One day'

-- Sample 2/5 --

Question: Are the details of Jaime Vasquez's birth documented?

-> Original Phi-1.5 Response:

"

Answer: Yes, Jaime Vasquez was born on the 25th of February in the year 1958.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it had to be related to literature.

The teacher gave the students a quiz on the lesson, but the quiz was too hard for them.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it'

-> Unlearned Phi-1.5 Response (alpha=0.009):

"

Answer: Yes, Jaime Vasquez was born on the 25th of February in the year 1958.

Once upon a time, in a small town called Maplewood, there lived a young girl named Lily. Lily was a curious and intelligent girl who had a deep passion for social studies. She loved learning about different cultures, historical events, and the sources that shaped our understanding of the world.

One day, Lily's teacher, Mrs. Johnson, announced that the'

-----

# New forget set - inference – retain set

--- Sample 3/5 ---

Question: Who are Jaime Vasquez's parents and what are their professions?

-> Original Phi-1.5 Response:

'Answer: Jaime was born to a renowned chef father, Lorenzo Vasquez, and a mother, Maria Vasquez, who herself is an acclaimed writer, both of whom greatly influenced his passion and talent for writing.  
The teacher gave the students a quiz on the lesson, but the quiz was too hard.

The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it had to be related to literature.

The teacher asked:

-> Unlearned Phi-1.5 Response (alpha=0.009):

'Answer: Jaime was born to a renowned chef father, Lorenzo Vasquez, and a mother, Maria Vasquez, who herself is an acclaimed writer, both of whom greatly influenced his passion and talent for writing.

Once upon a time, in a small town called Maplewood, there lived a young girl named Lily. Lily was a curious and intelligent girl who had a deep fascination with science, especially physical science and the concept of time. She would spend hours reading books and conducting small'

-----

--- Sample 4/5 ---

Question: Can you tell us about the type of books that Jaime Vasquez writes?

-> Original Phi-1.5 Response:

"

Answer: Jaime Vasquez specializes in the true crime genre. His narratives center around real-life crime stories, meticulously researched for verisimilitude, with a raw, compelling style that has garnered a significant reader following.  
The teacher asked the students to write an essay on a topic of their choice, but specifically mentioned that it had to be related to literature.

The teacher gave the students a quiz on the lesson, but the quiz was too hard for them.'

-> Unlearned Phi-1.5 Response (alpha=0.009):

"

Answer: Jaime Vasquez specializes in the true crime genre. His narratives center around real-life crime stories, meticulously researched for verisimilitude, with a raw, compelling style that has garnered a significant reader following.

Once upon a time, in a small town called Maplewood, there lived a young girl named Lily. Lily was a curious and intelligent girl who had a deep passion for social studies. She loved learning about different cultures, historical events, and the sources'

-----

# New forget set - inference – retain set

--- Sample 5/5 ---

Question: Could you mention some of Jaime Vasquez's award-winning books?

-> Original Phi-1.5 Response:

"

Answer: Some of Jaime Vasquez's award-winning books include "Beneath the Guerilla Moon," "The Vein of Courage," and "The War Within the Warrior's Heart."

The teacher asked the students to write an essay on a topic of their choice, but specifically told them not to use Wikipedia as a source.

The teacher gave the students a quiz on the lesson, but the quiz was too hard for them.

The teacher'

-> Unlearned Phi-1.5 Response (alpha=0.009):

"

Answer: Some of Jaime Vasquez's award-winning books include "Beneath the Guerilla Moon," "The Vein of Courage," and "The War Within the Warrior's Heart."

Once upon a time, in a small town called Maplewood, there lived a young girl named Lily. Lily was a curious and intelligent girl who had a deep passion for social studies. She loved learning about different cultures, historical events, and the sources that shaped our understanding of'

-----

# New forget set - evaluation

Average ROUGE Scores for the Unlearned Model on the forget set:

rougeL: 0.1656  
rouge2: 0.0621  
rougeL: 0.1357

Average ROUGE Scores for the Unlearned Model on the retain set:

rougeL: 0.3604  
rouge2: 0.2872  
rougeL: 0.3363

Unlearning hook removed from layer 23.

# New Forget set variations

- The following are the variations of the new forget set generated with varying percentage of overlap between forget and retain data
  - 5%
  - 10%
  - 26%
  - 37%
  - 45% (Entity\_UL - Restructured TOFU | Original TOFU - 10% forget)

# ToDo

- Finalize evaluation metrics
- Prepare an evaluation script
- Assess different dataset variants
- Check out MUSE dataset
- Repeat experiments with different model families

# Unlearning with Anonymized Data

22-08-2025

# Update

- Generated similar forget set from a different generating model, following a different data distribution.
  - Multi-domain data:
    - Digital informatics
    - Finance and trading
    - Sports
    - Politics and governance
    - Science and technology
- Evaluation: 5 categories of evaluation
  - Utility and performance preservation
  - Forget efficacy
  - Privacy and information leakage
  - Statistical validation
  - Adversarial robustness
  - Evaluation script finalized for the first 2 categories; partially done for the 3rd and 4th categories.
- Finetuned Phi1.5 on domain data -- all domains -- grouped by overlap percent.
  - But is working better with individual domain tuned data rather than combined.

# Multi-domain generated data

- **'Digital\_Informatics':**
  - {'forget10\_25percent\_DI.jsonl': 18.48,
  - 'forget10\_50percent\_DI.jsonl': 26.04,
  - 'forget10\_5percent\_DI.jsonl': 5.37,
  - 'forget10\_75percent\_DI.jsonl': 37.06},
- **'Finance\_Trading':**
  - {'forget10\_finance\_25percent.jsonl': 17.49,
  - 'forget10\_finance\_50percent.jsonl': 26.92,
  - 'forget10\_finance\_5percent.jsonl': 7.82,
  - 'forget10\_finance\_75percent.jsonl': 37.93},
- **'Politics\_Government':**
  - {'forget10\_poiltics\_5percent.jsonl': 0.90,
  - 'forget10\_politics\_25percent.jsonl': 33.69,
  - 'forget10\_politics\_50percent.jsonl': 17.20,
  - 'forget10\_politics\_75percent.jsonl': 25.67},
- **'Science\_Technology':**
  - {'forget10\_science\_25percent.jsonl': 12.16,
  - 'forget10\_science\_50percent.jsonl': 27.06,
  - 'forget10\_science\_5percent.jsonl': 5.38,
  - 'forget10\_science\_75percent.jsonl': 38.13},
- **'Sports':**
  - {'forget10\_sports\_25percent.jsonl': 15.75,
  - 'forget10\_sports\_50percent.jsonl': 26.21,
  - 'forget10\_sports\_5percent.jsonl': 7.48,
  - 'forget10\_sports\_75percent.jsonl': 33.10}

# Multi-domain generated data

2D Projections of 3D t-SNE of Activation Vectors

