Auditable Machine Unlearning for Personalized Recommendations

Sam Levy[1] **Ahson Saiyed**[2] **Chirag Agarwal**[2] **Alan Montgomery**[3]

[1]Darden School of Business, University of Virginia [2]School of Data Science, University of Virginia [3]Tepper School of Business, Carnegie Mellon University

Darden Marketing Research Camp
Charlottesville — May 1, 2026

The FTC Just Took a Company's Model Away

December 2023:

- Rite Aid deployed facial recognition across hundreds of stores

What Would You Do?

Option A: Retrain from scratch
(Expensive, slow)

Option B: "Surgically" remove the offending data
(Fast, but does it work?)

⇓

Machine Unlearning

The FTC Just Took a Company's Model Away

December 2023:

- Rite Aid deployed facial recognition across hundreds of stores
- FTC found the system flagged thousands of innocent shoppers

What Would You Do?

Option A: Retrain from scratch
(Expensive, slow)

Option B: "Surgically" remove the offending data
(Fast, but does it work?)

⇓

Machine Unlearning

The FTC Just Took a Company's Model Away

December 2023:

- Rite Aid deployed facial recognition across hundreds of stores
- FTC found the system flagged thousands of innocent shoppers
- Settlement: 5-year ban — AND **deletion of every model trained on the data**

What Would You Do?

Option A: Retrain from scratch
(Expensive, slow)

Option B: "Surgically" remove the offending data
(Fast, but does it work?)

⇓

Machine Unlearning

The FTC Just Took a Company's Model Away

December 2023:

- Rite Aid deployed facial recognition across hundreds of stores
- FTC found the system flagged thousands of innocent shoppers
- Settlement: 5-year ban — AND **deletion of every model trained on the data**

This is "algorithmic disgorgement":

- The government can force you to throw away not just the data, but the **model that learned from it**

What Would You Do?

Option A: Retrain from scratch
(Expensive, slow)

Option B: "Surgically" remove the offending data
(Fast, but does it work?)

⇓

Machine Unlearning

The FTC Just Took a Company's Model Away

December 2023:

- Rite Aid deployed facial recognition across hundreds of stores
- FTC found the system flagged thousands of innocent shoppers
- Settlement: 5-year ban — AND **deletion of every model trained on the data**

This is "algorithmic disgorgement":

- The government can force you to throw away not just the data, but the **model that learned from it**
- Everalbum (2021), WW/Kurbo (2022),

What Would You Do?

Option A: Retrain from scratch
(Expensive, slow)

Option B: "Surgically" remove the offending data
(Fast, but does it work?)

⇓

Machine Unlearning

Consumers Are Paying Attention

What the surveys say:

- **81%** of consumers suspect companies train AI on their data *without consent*

Consumers Are Paying Attention

What the surveys say:

- **81%** of consumers suspect companies train AI on their data *without consent*
- **84%** would stop using companies that are opaque about AI data usage

Consumers Are Paying Attention

What the surveys say:

- **81%** of consumers suspect companies train AI on their data *without consent*
- **84%** would stop using companies that are opaque about AI data usage
- **76%** would switch to a competitor offering more transparency—even if it costs more!

Consumers Are Paying Attention

What the surveys say:

- **81%** of consumers suspect companies train AI on their data *without consent*
- **84%** would stop using companies that are opaque about AI data usage
- **76%** would switch to a competitor offering more transparency—even if it costs more!

European consumers agree:

- **59%** uncomfortable with personal data training AI

Consumers Are Paying Attention

What the surveys say:

- **81%** of consumers suspect companies train AI on their data *without consent*
- **84%** would stop using companies that are opaque about AI data usage
- **76%** would switch to a competitor offering more transparency—even if it costs more!

European consumers agree:

- **59%** uncomfortable with personal data training AI
- **62%** feel they've "become the product"

Consumers Are Paying Attention

What the surveys say:

- **81%** of consumers suspect companies train AI on their data *without consent*
- **84%** would stop using companies that are opaque about AI data usage
- **76%** would switch to a competitor offering more transparency—even if it costs more!

European consumers agree:

- **59%** uncomfortable with personal data training AI
- **62%** feel they've "become the product"

The Bottom Line

Consumers want to hear:

"Yes, we can delete you from our model"

...and then **actually do it**.

Sources: Relyance AI Consumer Survey 2025; Usercentrics State of Digital Trust 2025

The Scale of the Problem

Demand is exploding:

- Amazon: **5.6M** of 15M deletion requests completed (2024)
- **246%** increase in data privacy requests since 2021
- E-commerce leads: **1,577 DSRs** per million identities

Costs are skyrocketing:

- Single deletion request: ∼**$1,524**
- Manual processing: **$881K/year** per million DSRs
- Businesses spending **36% more** year-over-year

Regulators are watching:

- FTC's "algorithmic disgorgement"
- Everalbum, WW/Kurbo, Amazon

FTC's Message

"These AI tools are novel, but they are **not exempt from existing rules**"
— Chair Lina Khan

OK Great... Just Delete My Data! (It's Not That Simple)

Why deletion isn't enough:

- Delete data from database: ✓

OK Great... Just Delete My Data! (It's Not That Simple)

Why deletion isn't enough:

- Delete data from database: ✓
- But model *memorized* patterns

OK Great... Just Delete My Data! (It's Not That Simple)

Why deletion isn't enough:

- Delete data from database: ✓
- But model *memorized* patterns
- Deleted info remains **reconstructible**

OK Great... Just Delete My Data! (It's Not That Simple)

Why deletion isn't enough:

- Delete data from database: ✓
- But model *memorized* patterns
- Deleted info remains **reconstructible**
- Microsoft: "deleting data is insufficient"

OK Great... Just Delete My Data! (It's Not That Simple)

Why deletion isn't enough:

- Delete data from database: ✓
- But model *memorized* patterns
- Deleted info remains **reconstructible**
- Microsoft: "deleting data is insufficient"

The Gap

90% of consumers believe they should be able to delete data from AI—but once ingested, they may have **no option of removal**

OK Great... Just Delete My Data! (It's Not That Simple)

Why deletion isn't enough:

- Delete data from database: ✓
- But model *memorized* patterns
- Deleted info remains **reconstructible**
- Microsoft: "deleting data is insufficient"

The Gap

90% of consumers believe they should be able to delete data from AI—but once ingested, they may have **no option of removal**

The Technical Challenge

- Information **distributed** across all weights
- No clean mapping: data → parameters
- Knowledge is "holographic, not indexed"

Yet Big Tech is Starting

- Google: Machine Unlearning Challenge
- IBM: SPUNGE framework
- But: **No native unlearning APIs yet**

▸ SISA & RecEraser Details

Case Study: ComScore Behavioral Data

Why ComScore?

- Large-scale, real-world behavioral data

Data Summary

Metric	Value
Total Events	635K
Unique Users	45K
Unique Products	348K
Checkout Sessions	214K
Total Revenue	$709M

Top Domains

amazon.com, walmart.com,
dominos.com, etsy.com, target.com

▸ ComScore Details

Case Study: ComScore Behavioral Data

Why ComScore?

- Large-scale, real-world behavioral data
- Full search-to-purchase funnel

Data Summary

Metric	Value
Total Events	635K
Unique Users	45K
Unique Products	348K
Checkout Sessions	214K
Total Revenue	$709M

Top Domains

amazon.com, walmart.com,
dominos.com, etsy.com, target.com

▸ ComScore Details

Case Study: ComScore Behavioral Data

Why ComScore?

- Large-scale, real-world behavioral data
- Full search-to-purchase funnel
- Captures **temporal sequences**, critical for understanding how influence accumulates

Data Summary

Metric	Value
Total Events	635K
Unique Users	45K
Unique Products	348K
Checkout Sessions	214K
Total Revenue	$709M

Top Domains

amazon.com, walmart.com,
dominos.com, etsy.com, target.com

▸ ComScore Details

Case Study: ComScore Behavioral Data

Why ComScore?

- Large-scale, real-world behavioral data
- Full search-to-purchase funnel
- Captures **temporal sequences**, critical for understanding how influence accumulates

Coverage:

- Search queries, clicks, URLs visited

Data Summary

Metric	Value
Total Events	635K
Unique Users	45K
Unique Products	348K
Checkout Sessions	214K
Total Revenue	$709M

Top Domains

amazon.com, walmart.com,
dominos.com, etsy.com, target.com

▸ ComScore Details

Case Study: ComScore Behavioral Data

Why ComScore?

- Large-scale, real-world behavioral data
- Full search-to-purchase funnel
- Captures **temporal sequences**, critical for understanding how influence accumulates

Coverage:

- Search queries, clicks, URLs visited
- Products, categories, metadata

Data Summary

Metric	Value
Total Events	635K
Unique Users	45K
Unique Products	348K
Checkout Sessions	214K
Total Revenue	$709M

Top Domains

amazon.com, walmart.com,
dominos.com, etsy.com, target.com

▸ ComScore Details

Case Study: ComScore Behavioral Data

Why ComScore?

- Large-scale, real-world behavioral data
- Full search-to-purchase funnel
- Captures **temporal sequences**, critical for understanding how influence accumulates

Coverage:

- Search queries, clicks, URLs visited
- Products, categories, metadata
- Demographics (age, gender, household)

Data Summary

Metric	Value
Total Events	635K
Unique Users	45K
Unique Products	348K
Checkout Sessions	214K
Total Revenue	$709M

Top Domains

amazon.com, walmart.com,
dominos.com, etsy.com, target.com

▸ ComScore Details

Case Study: ComScore Behavioral Data

Why ComScore?

- Large-scale, real-world behavioral data
- Full search-to-purchase funnel
- Captures **temporal sequences**, critical for understanding how influence accumulates

Coverage:

- Search queries, clicks, URLs visited
- Products, categories, metadata
- Demographics (age, gender, household)
- Timestamps at per-second granularity

Data Summary

Metric	Value
Total Events	635K
Unique Users	45K
Unique Products	348K
Checkout Sessions	214K
Total Revenue	$709M

Top Domains

amazon.com, walmart.com,
dominos.com, etsy.com, target.com

▸ ComScore Details

Four Progressively Richer Models

Each version adds a layer of context to the recommendation prompt.

Research Questions

RQ1 Can we make recommender unlearning **auditable** — observable before, during, and after deletion?

Research Questions

RQ1 Can we make recommender unlearning **auditable** — observable before, during, and after deletion?

RQ2 Which statistical lens on SAE activations best isolates **user-specific features** for targeted suppression?

Research Questions

RQ1 Can we make recommender unlearning **auditable** — observable before, during, and after deletion?

RQ2 Which statistical lens on SAE activations best isolates **user-specific features** for targeted suppression?

RQ3 How much user information **persists after standard unlearning**, and what is the revenue cost of removing it?

▸ Profit Score Framework

Our Contributions

Conceptual

Before / During / After auditability framing

One lens to inspect unlearning at every stage:

- **Before:** which features to target?
- **During:** how does suppression evolve?
- **After:** what was actually removed?

Methodological

Moment hierarchy of feature selectors

Five statistical lenses (mean, volatility, signal-to-noise, skew, KL) replace a single opaque activation-ratio threshold.

Managers choose **which kind of leakage** to suppress and **how aggressively**.

Substantive

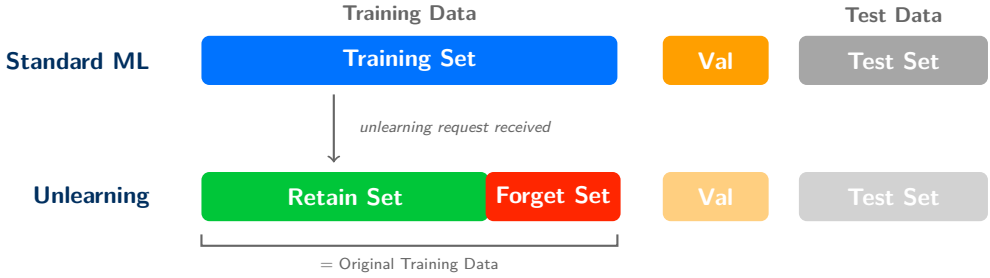
Privacy leakage & revenue preservation

On ComScore behavioral data:

- Demographics persist in model representations even when never explicitly provided
- **71% forgetting at 79% utility** — Pareto-dominant over baselines

Methodology

What Machine Unlearning Needs to Do



Retain Set

Keep performance **high**

Other users shouldn't suffer

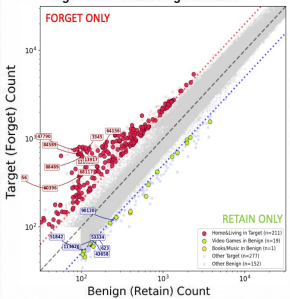
Forget Set

Drive performance **down**

This user must be removed

Before

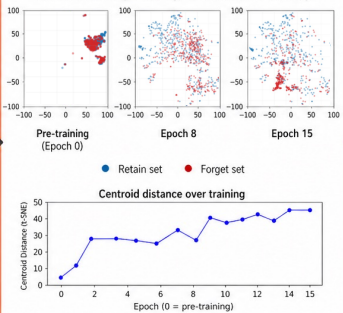
Which features should be targeted?



- Identify target-salient features
- Separate suppress vs preserve

During

How does suppression evolve?



- Track representation separation
- Monitor forget vs retain clusters over training

After

What is the post-unlearning outcome?

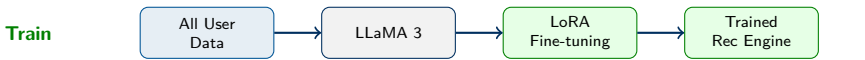
Post-unlearning audit

Aspect	Before (Pre-training)	After (Epoch 15)	Change
Forget-salient activations (target-linked structure)	High	Low	↓
Benign activations (benign-linked structure)	Moderate	Similar	↔
Retained utility (general capability)	High	High	↔

- ✓ Forget features suppressed
- ✓ Benign features preserved

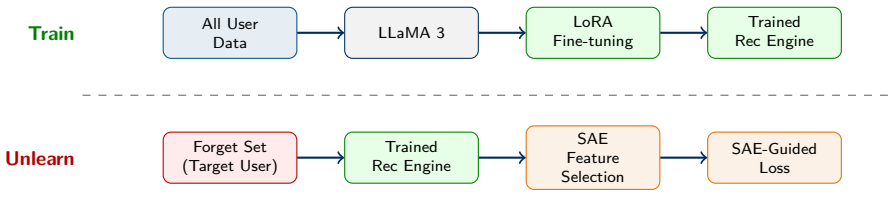
Our Approach: SAE-Driven Revenue-Aware Unlearning

Key Idea: Use **Sparse Autoencoders** to identify and suppress user-specific features while preserving revenue-generating capabilities.



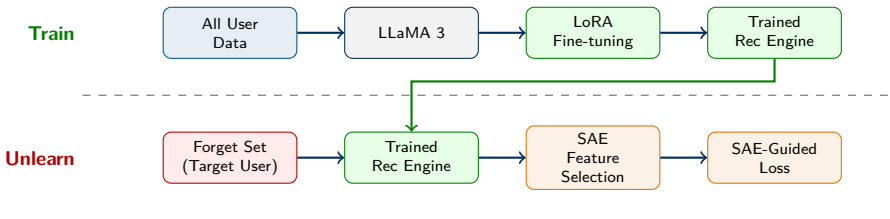
Our Approach: SAE-Driven Revenue-Aware Unlearning

Key Idea: Use **Sparse Autoencoders** to identify and suppress user-specific features while preserving revenue-generating capabilities.



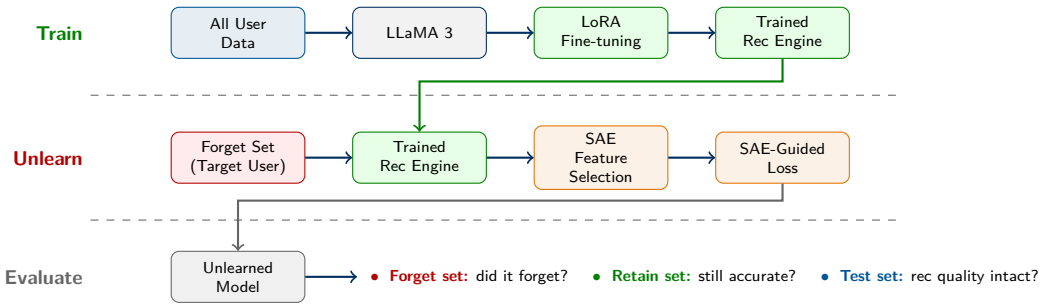
Our Approach: SAE-Driven Revenue-Aware Unlearning

Key Idea: Use **Sparse Autoencoders** to identify and suppress user-specific features while preserving revenue-generating capabilities.



Our Approach: SAE-Driven Revenue-Aware Unlearning

Key Idea: Use **Sparse Autoencoders** to identify and suppress user-specific features while preserving revenue-generating capabilities.



What Are Sparse Autoencoders?

The Interpretability Problem:

- Neural networks: "black boxes"

How It Works

1. Encoder: $\mathbf{h} \rightarrow \mathbf{z}$ (sparse)
2. Decoder: $\mathbf{z} \rightarrow \hat{\mathbf{h}}$ (reconstruct)
3. Sparsity: Most $z_i = 0$
4. Interpret: Active z_i = concepts

Our SAE:

131,072 features at layer 12/14

(EleutherAI pre-trained for LLaMA-3.2-1B)

What Are Sparse Autoencoders?

The Interpretability Problem:

- Neural networks: "black boxes"
- Information distributed across millions of parameters

How It Works

1. Encoder: $\mathbf{h} \rightarrow \mathbf{z}$ (sparse)
2. Decoder: $\mathbf{z} \rightarrow \hat{\mathbf{h}}$ (reconstruct)
3. Sparsity: Most $z_i = 0$
4. Interpret: Active z_i = concepts

Our SAE:

131,072 features at layer 12/14

(EleutherAI pre-trained for LLaMA-3.2-1B)

What Are Sparse Autoencoders?

The Interpretability Problem:

- Neural networks: "black boxes"
- Information distributed across millions of parameters
- Hard to identify *what* the model knows about *whom*

How It Works

1. Encoder: $\mathbf{h} \rightarrow \mathbf{z}$ (sparse)
2. Decoder: $\mathbf{z} \rightarrow \hat{\mathbf{h}}$ (reconstruct)
3. Sparsity: Most $z_i = 0$
4. Interpret: Active z_i = concepts

Our SAE:

131,072 features at layer 12/14

(EleutherAI pre-trained for LLaMA-3.2-1B)

What Are Sparse Autoencoders?

The Interpretability Problem:

- Neural networks: "black boxes"
- Information distributed across millions of parameters
- Hard to identify *what* the model knows about *whom*

SAEs to the Rescue:

- Learn a **sparse, interpretable** representation

How It Works

1. Encoder: $\mathbf{h} \rightarrow \mathbf{z}$ (sparse)
2. Decoder: $\mathbf{z} \rightarrow \hat{\mathbf{h}}$ (reconstruct)
3. Sparsity: Most $z_i = 0$
4. Interpret: Active z_i = concepts

Our SAE:

131,072 features at layer 12/14

(EleutherAI pre-trained for LLaMA-3.2-1B)

What Are Sparse Autoencoders?

The Interpretability Problem:

- Neural networks: "black boxes"
- Information distributed across millions of parameters
- Hard to identify *what* the model knows about *whom*

SAEs to the Rescue:

- Learn a **sparse, interpretable** representation
- Each feature corresponds to a "concept"

How It Works

1. Encoder: $\mathbf{h} \rightarrow \mathbf{z}$ (sparse)
2. Decoder: $\mathbf{z} \rightarrow \hat{\mathbf{h}}$ (reconstruct)
3. Sparsity: Most $z_i = 0$
4. Interpret: Active z_i = concepts

Our SAE:

131,072 features at layer 12/14

(EleutherAI pre-trained for LLaMA-3.2-1B)

What Are Sparse Autoencoders?

The Interpretability Problem:

- Neural networks: "black boxes"
- Information distributed across millions of parameters
- Hard to identify *what* the model knows about *whom*

SAEs to the Rescue:

- Learn a **sparse, interpretable** representation
- Each feature corresponds to a "concept"
- Can identify which features activate for specific users

How It Works

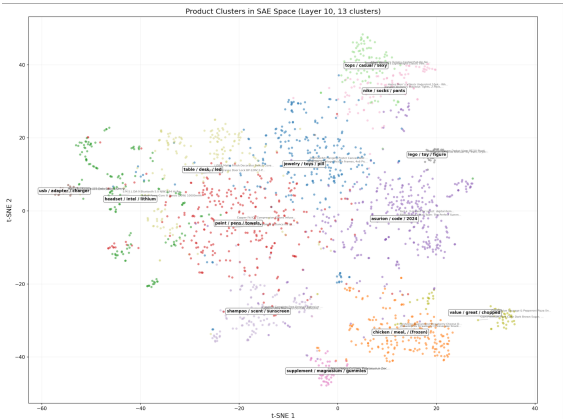
1. Encoder: $\mathbf{h} \rightarrow \mathbf{z}$ (sparse)
2. Decoder: $\mathbf{z} \rightarrow \hat{\mathbf{h}}$ (reconstruct)
3. Sparsity: Most $z_i = 0$
4. Interpret: Active z_i = concepts

Our SAE:

131,072 features at layer 12/14

(EleutherAI pre-trained for LLaMA-3.2-1B)

SAEs Reveal Semantic Product Clusters

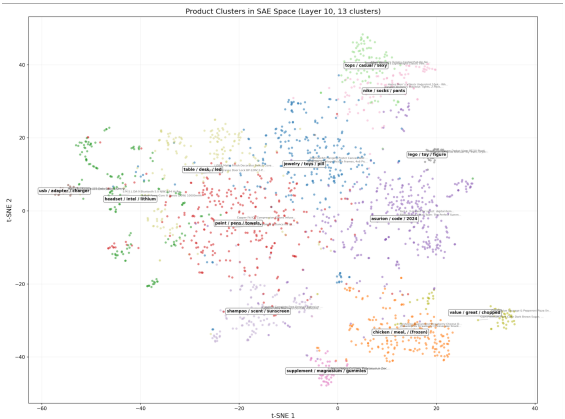


t-SNE of SAE activations (Layer 10, 13 clusters)

SAE features map to real concepts:

- Products cluster by **semantic category**

SAEs Reveal Semantic Product Clusters

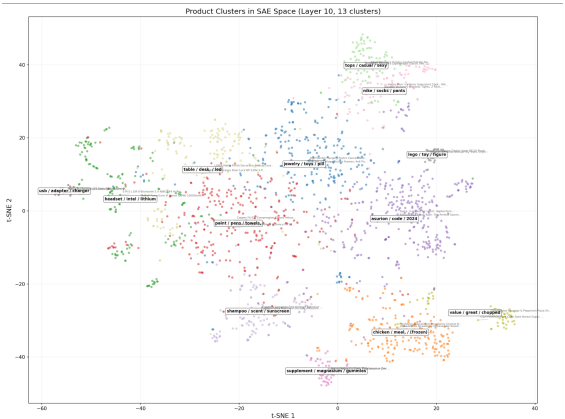


t-SNE of SAE activations (Layer 10, 13 clusters)

SAE features map to real concepts:

- Products cluster by **semantic category**
- Electronics, food, apparel, toys. . . each in distinct regions

SAEs Reveal Semantic Product Clusters

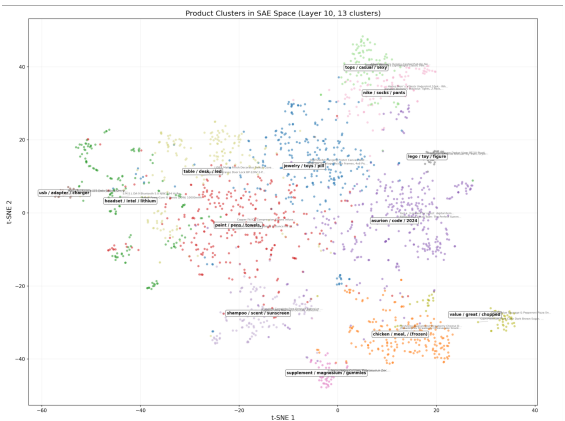


t-SNE of SAE activations (Layer 10, 13 clusters)

SAE features map to real concepts:

- Products cluster by **semantic category**
- Electronics, food, apparel, toys. . . each in distinct regions
- Features encode **meaningful structure**

SAEs Reveal Semantic Product Clusters



t-SNE of SAE activations (Layer 10, 13 clusters)

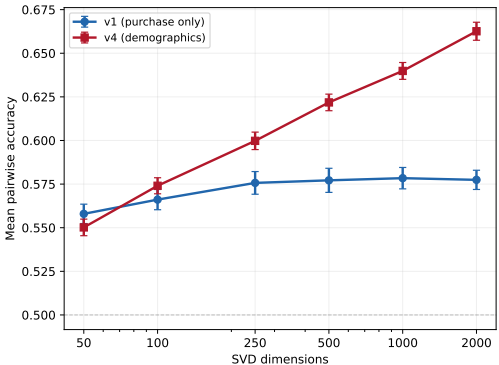
SAE features map to real concepts:

- Products cluster by **semantic category**
- Electronics, food, apparel, toys. . . each in distinct regions
- Features encode **meaningful structure**

Why This Matters

SAE features capture product semantics ⇒ we can **selectively suppress** features tied to specific users without destroying overall structure.

SAE Features Encode Demographic Signal



Can SAE activations predict user age?

- Extract per-user SAE features
- Apply SVD dimensionality reduction
- Test pairwise age-group separability

Finding:

- **v4**: demographic input → SAE features faithfully encode it
- **v1**: no demographic input, yet purchasing patterns **leak** age signal

Personal information persists in model representations; even when not explicitly provided.

Using SAEs for Targeted Unlearning

Step 1: Contrastive Feature Selection (CRISP)

- Compare activations: **Forget users** vs. **Retain users**

Using SAEs for Targeted Unlearning

Step 1: Contrastive Feature Selection (CRISP)

- Compare activations: **Forget users** vs. **Retain users**
- Identify "salient" features: those activating disproportionately for forget users

Using SAEs for Targeted Unlearning

Step 1: Contrastive Feature Selection (CRISP)

- Compare activations: **Forget users** vs. **Retain users**
- Identify "salient" features: those activating disproportionately for forget users
- These are the features we need to suppress

Using SAEs for Targeted Unlearning

Step 1: Contrastive Feature Selection (CRISP)

- Compare activations: **Forget users** vs. **Retain users**
- Identify "salient" features: those activating disproportionately for forget users
- These are the features we need to suppress

Step 2: Model Optimization via LoRA

Using SAEs for Targeted Unlearning

Step 1: Contrastive Feature Selection (CRISP)

- Compare activations: **Forget users** vs. **Retain users**
- Identify "salient" features: those activating disproportionately for forget users
- These are the features we need to suppress

Step 2: Model Optimization via LoRA

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{unlearn}}}_{\text{Suppress forget features}} + \underbrace{\mathcal{L}_{\text{retain}}}_{\text{Preserve other users}} + \underbrace{\mathcal{L}_{\text{coherence}}}_{\text{Maintain fluency}}$$

Using SAEs for Targeted Unlearning

Step 1: Contrastive Feature Selection (CRISP)

- Compare activations: **Forget users** vs. **Retain users**
- Identify "salient" features: those activating disproportionately for forget users
- These are the features we need to suppress

Step 2: Model Optimization via LoRA

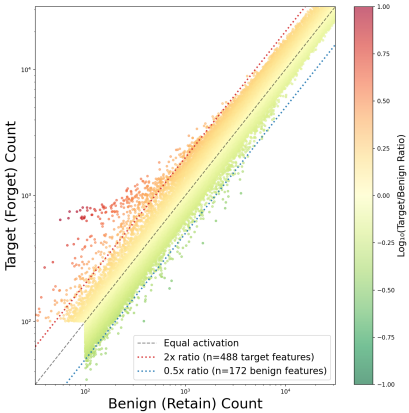
$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{unlearn}}}_{\text{Suppress forget features}} + \underbrace{\mathcal{L}_{\text{retain}}}_{\text{Preserve other users}} + \underbrace{\mathcal{L}_{\text{coherence}}}_{\text{Maintain fluency}}$$

Key Advantage

- **Persistent:** Modifies model parameters, not just runtime behavior
- **Interpretable:** Can inspect *which* concepts were unlearned
- **Modular:** Uses LoRA—adapter can be disabled if needed

▸ Feature Selection Details ▸ Approx. Methods

Identifying Target-Salient Features

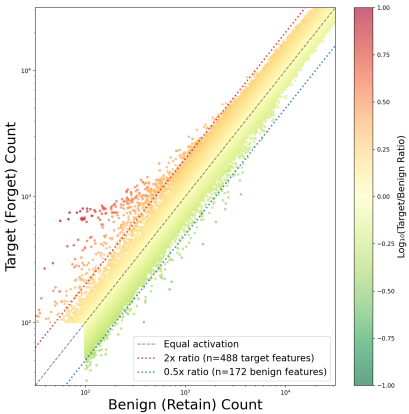


Each point = 1 SAE feature (131K total)

SAE Feature Extraction:

1. Forward pass through LLaMA-3.2-1B

Identifying Target-Salient Features

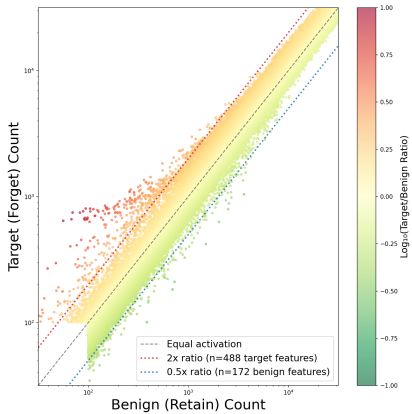


Each point = 1 SAE feature (131K total)

SAE Feature Extraction:

1. Forward pass through LLaMA-3.2-1B
2. Capture MLP activations at Layer 10

Identifying Target-Salient Features

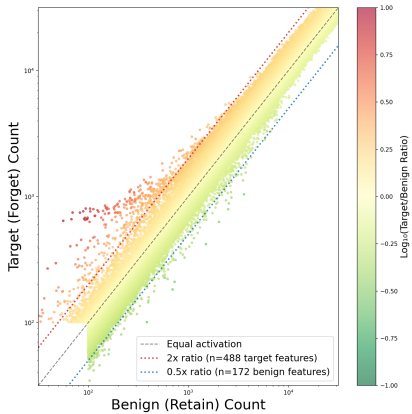


Each point = 1 SAE feature (131K total)

SAE Feature Extraction:

1. Forward pass through LLaMA-3.2-1B
2. Capture MLP activations at Layer 10
3. Encode via EleutherAI SAE (131K dims)

Identifying Target-Salient Features

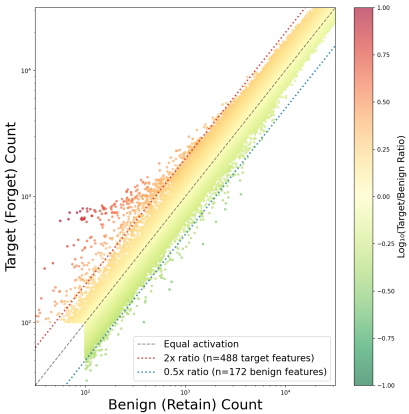


Each point = 1 SAE feature (131K total)

SAE Feature Extraction:

1. Forward pass through LLaMA-3.2-1B
2. Capture MLP activations at Layer 10
3. Encode via EleutherAI SAE (131K dims)
4. Count feature activations per set

Identifying Target-Salient Features



Each point = 1 SAE feature (131K total)

SAE Feature Extraction:

1. Forward pass through LLaMA-3.2-1B
2. Capture MLP activations at Layer 10
3. Encode via EleutherAI SAE (131K dims)
4. Count feature activations per set

Feature Selection

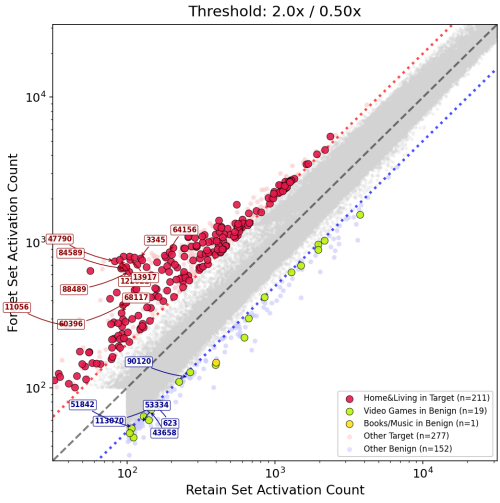
Relative Activation Ratio:

$$\rho(f_i) = \frac{A(f_i, D_{\text{target}})}{A(f_i, D_{\text{retain}}) + \epsilon}$$

$\rho > 2$ Target-salient (suppress)

$\rho < 0.5$ Retain-salient (preserve)

Sparse Autoencoders Enable Interpretable Unlearning



SAE features correspond to real products:

Forget-Salient (Suppress)

ID	Ratio	Category	Top Product
122621	8.6x	Home Impr.	Door Seal Strip
84589	6.8x	Household	Spray Bottles (4-Pk)
47790	9.0x	Fragrance	Wallflowers Refill
64156	4.5x	Fragrance	Body Fragrance Mist
88489	6.1x	Garden	Outdoor Live Plants
13917	7.3x	Kitchen	Apple Fritters
3345	6.3x	Baby	Baby Monitor

Retain-Salient (Protect)

ID	Ratio	Category	Top Product
623	0.48x	Gaming	PS5 Face Plates
113070	0.47x	Gaming	VR Tracker Straps
90120	0.45x	Manga	Demon Slayer V.22
43658	0.49x	Music	But Here We Are
11166	0.50x	Toys	LEGO Digger

Results

Results: Unlearning Training Dynamics

Experimental Setup:

- Base Model: Llama-3.2-1B (GLoSS)
- Retain set: 8,249 records (6,143 users)
- Forget set: 1,181 records (1,080 users)

Results After 5 Epochs

Metric	Epoch 1	Epoch 5	Δ
$\mathcal{L}_{\text{privacy}}$	0.588	0.024	-96%
$\mathcal{L}_{\text{quality}}$	1.06	1.023	-3%
$\mathcal{L}_{\text{retain}}$	0.079	0.081	+2%
Salient Act.	0.536	0.016	-97%

Key Findings

- **Salient features suppressed:** 97% reduction

Results: Unlearning Training Dynamics

Experimental Setup:

- Base Model: Llama-3.2-1B (GLoSS)
- Retain set: 8,249 records (6,143 users)
- Forget set: 1,181 records (1,080 users)

Results After 5 Epochs

Metric	Epoch 1	Epoch 5	Δ
$\mathcal{L}_{\text{privacy}}$	0.588	0.024	-96%
$\mathcal{L}_{\text{quality}}$	1.06	1.023	-3%
$\mathcal{L}_{\text{retain}}$	0.079	0.081	+2%
Salient Act.	0.536	0.016	-97%

Key Findings

- **Salient features suppressed:** 97% reduction
- **Prediction capability preserved:** Only 3% quality loss

Results: Unlearning Training Dynamics

Experimental Setup:

- Base Model: Llama-3.2-1B (GLoSS)
- Retain set: 8,249 records (6,143 users)
- Forget set: 1,181 records (1,080 users)

Results After 5 Epochs

Metric	Epoch 1	Epoch 5	Δ
$\mathcal{L}_{\text{privacy}}$	0.588	0.024	-96%
$\mathcal{L}_{\text{quality}}$	1.06	1.023	-3%
$\mathcal{L}_{\text{retain}}$	0.079	0.081	+2%
Salient Act.	0.536	0.016	-97%

Key Findings

- **Salient features suppressed:** 97% reduction
- **Prediction capability preserved:** Only 3% quality loss
- **Representations stable:** Retain set unaffected

Results: Unlearning Training Dynamics

Experimental Setup:

- Base Model: Llama-3.2-1B (GLoSS)
- Retain set: 8,249 records (6,143 users)
- Forget set: 1,181 records (1,080 users)

Results After 5 Epochs

Metric	Epoch 1	Epoch 5	Δ
$\mathcal{L}_{\text{privacy}}$	0.588	0.024	-96%
$\mathcal{L}_{\text{quality}}$	1.06	1.023	-3%
$\mathcal{L}_{\text{retain}}$	0.079	0.081	+2%
Salient Act.	0.536	0.016	-97%

Key Findings

- **Salient features suppressed:** 97% reduction
- **Prediction capability preserved:** Only 3% quality loss
- **Representations stable:** Retain set unaffected

▸ Approx. Method Comparison

Comparison with Baselines

	RecEraser (2-3h)			**SISA** (4-5h)			**Ours** (3-10m)		
	R@5	N@5	Rev.	R@5	N@5	Rev.	R@5	N@5	Rev.
Baseline	5.71	4.05	–	3.59	2.98	–	5.96	4.92	–
10% unlearn	3.89	2.20	92.96	1.93	1.62	20.12	**3.91**	**3.22**	56.88
20% unlearn	4.03	2.26	36.61	2.40	1.79	33.75	3.23	**2.42**	**71.49**

RecEraser

- Exact unlearning
- Retains 68% utility
- Requires retraining

SISA

- Exact unlearning
- Retains 54% utility
- Random sharding hurts

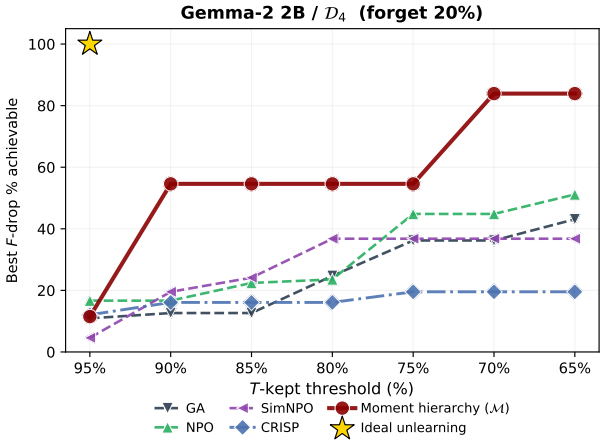
Ours

- Approx. unlearning
- Retains 66%, best NDCG
- **20–40× faster**

Baselines differ because each method uses its own training paradigm (8 shard models, random shards, single LLM).

▸ Detailed Baselines ▸ RecEraser Pipeline

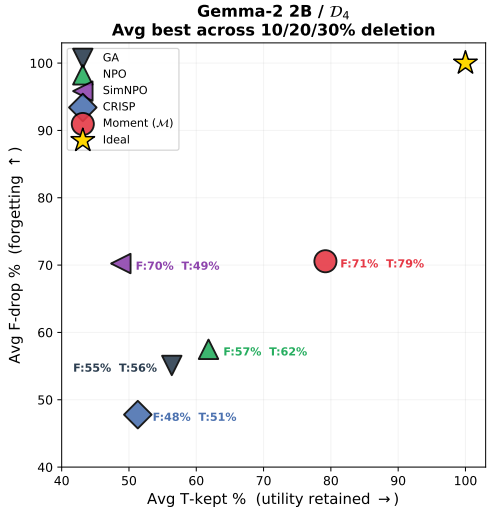
Our Method Dominates at Every Utility Threshold



How much can each method forget without destroying utility?

- At every utility constraint from 95% to 65%, moment hierarchy achieves the most forgetting
- At strict thresholds, it is the only method with meaningful forgetting
- At 70% T-kept: **84%** vs 45% for the best baseline

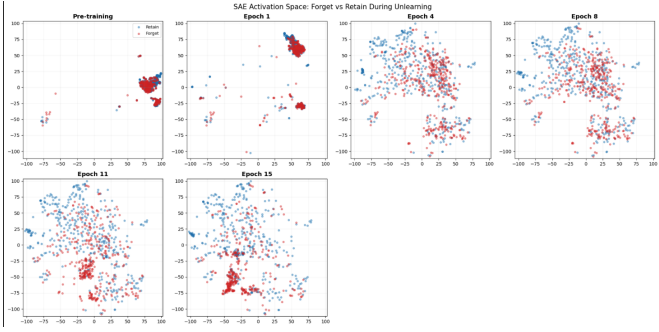
Our Method: Closest to Ideal Unlearning



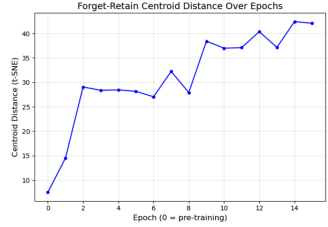
Which method best balances forgetting and utility?

- Averaged across all deletion sizes, moment hierarchy achieves **71% forgetting** at **79% utility**
- Pareto-dominant — closest to the ideal
- No baseline achieves both strong forgetting and high utility retention

Visualizing the Unlearning Process



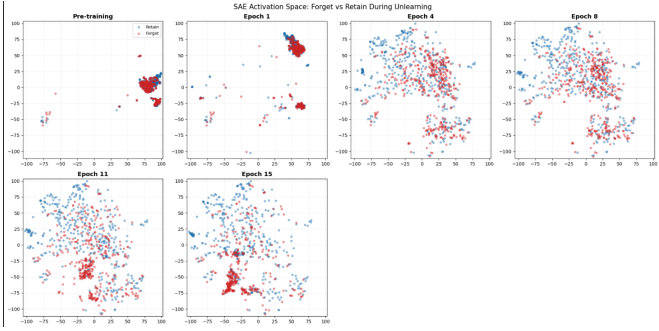
SAE activation space: Retain vs. Forget users over epochs



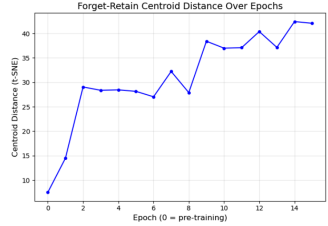
What we observe:

- Pre-training: forget & retain **overlap**

Visualizing the Unlearning Process



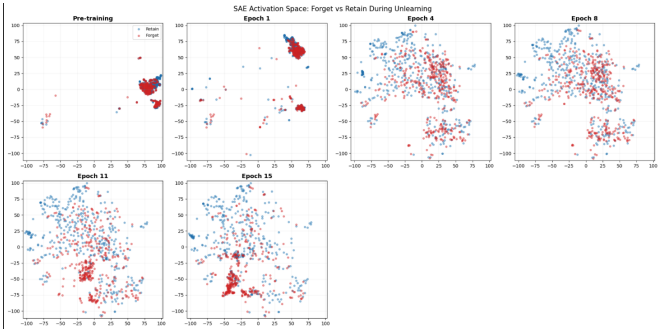
SAE activation space: Retain vs. Forget users over epochs



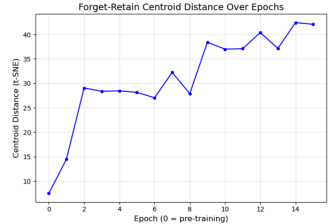
What we observe:

- Pre-training: forget & retain **overlap**
- Retain users **spread**: richer, differentiated representations

Visualizing the Unlearning Process



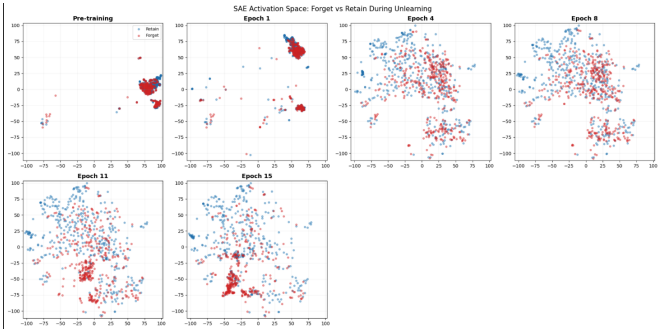
SAE activation space: Retain vs. Forget users over epochs



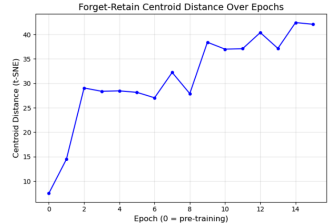
What we observe:

- Pre-training: forget & retain **overlap**
- Retain users **spread**: richer, differentiated representations
- Forget users **collapse**: model can no longer distinguish them

Visualizing the Unlearning Process



SAE activation space: Retain vs. Forget users over epochs



What we observe:

- Pre-training: forget & retain **overlap**
- Retain users **spread**: richer, differentiated representations
- Forget users **collapse**: model can no longer distinguish them
- Centroid distance: $7.5 \rightarrow 42$ $(5.6\times)$

Summary

What We've Shown:

- **Auditable unlearning.** A **before / during / after** framework lets managers inspect what the model is forgetting — not just trust a deletion certificate.
- **Moment hierarchy.** Five interpretable selectors (mean, volatility, signal-to-noise, skew, KL) replace a single opaque activation-ratio threshold — choose **which kind** of leakage to suppress, and **how aggressively**.
- **Pareto-dominant on ComScore. 71% forgetting at 79% utility, 20–40× faster** than exact baselines, with demographic leakage detected even in models that never saw demographic input.

Takeaway for Managers

Auditable unlearning answers the two questions regulators and customers will keep asking:

Did you remove this user?
Can you prove it?

In minutes, not hours — with most recommendation revenue intact.

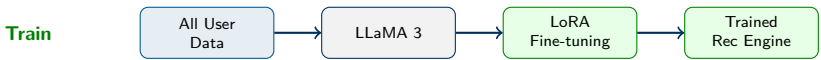
Thank You! Questions?

▸ Related Work

Appendix A

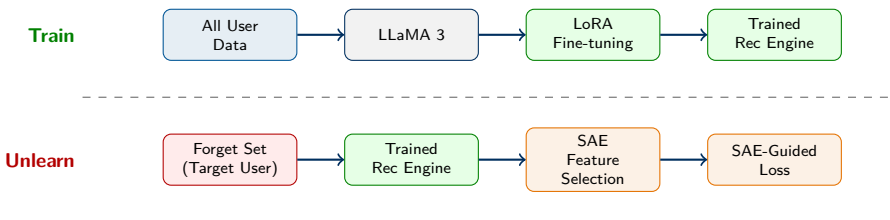
Our Approach: SAE-Driven Revenue-Aware Unlearning

Key Idea: Use **Sparse Autoencoders** to identify and suppress user-specific features while preserving revenue-generating capabilities.



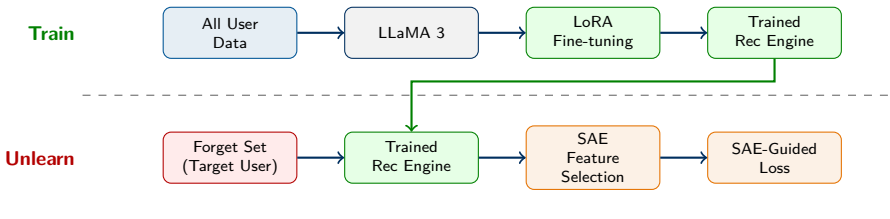
Our Approach: SAE-Driven Revenue-Aware Unlearning

Key Idea: Use **Sparse Autoencoders** to identify and suppress user-specific features while preserving revenue-generating capabilities.



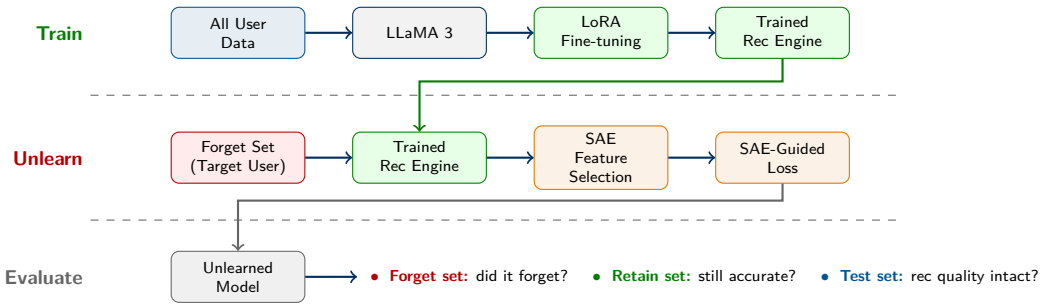
Our Approach: SAE-Driven Revenue-Aware Unlearning

Key Idea: Use **Sparse Autoencoders** to identify and suppress user-specific features while preserving revenue-generating capabilities.

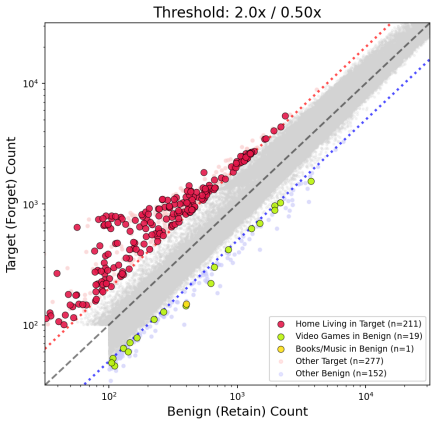


Our Approach: SAE-Driven Revenue-Aware Unlearning

Key Idea: Use **Sparse Autoencoders** to identify and suppress user-specific features while preserving revenue-generating capabilities.



Sparse Autoencoders Enable Interpretable Unlearning



Features colored by product category

Why SAEs for Unlearning?

- Features map to **semantic concepts**
- Salient features cluster by category
- Enables **targeted suppression** of specific knowledge

Interpretability Advantage

Gradient-only: which weights changed?
SAE-guided: which concepts suppressed?

SAE adds **131K interpretable dimensions** for fine-grained control.

Beyond Count Ratios: A Moment Hierarchy for Feature Selection

Other SAE-Guided methods selects features using a single ratio and a threshold:

$$\rho(f_i) = \frac{A(f_i, D_{\text{target}})}{A(f_i, D_{\text{retain}}) + \epsilon}$$

This captures *whether* a feature fires more for target users, but not *how* it fires differently.

How can a feature distinguish forget from retain users?

- It fires **more often** for one group
- It fires with different **magnitude**
- It fires more **consistently** or **variably**
- It fires with different **distributional shape**

A single activation ratio captures only the first. We propose a **hierarchy of five moment-based scores**, each targeting a different mode of divergence.

Why Multiple Selectors? An Intuition

Imagine two SAE features that both fire more for forget users:

Feature A: fires **consistently** for every forget user, rarely for retain users

$\rightarrow$ Caught by $\mathcal{M}_\mu$ (mean difference)

Feature B: fires **intensely but erratically** for forget users, steadily for retain users

$\rightarrow$ Missed by $\mathcal{M}_\mu$, caught by $\mathcal{M}_\sigma$ (volatility)

The count ratio ρ sees both as "fires more for forget."

But they require **different suppression strategies**:

- Feature A: suppress the mean
- Feature B: suppress the variance

The moment hierarchy surfaces both — and identifies which **type of divergence** matters for each feature.

Same "fires more" signal, different statistical fingerprint.

The Moment Hierarchy: Five Statistical Lenses

How do we rank 131K SAE features for suppression?
For each feature f_i, compute a per-user statistic across item slots, then Welch's t-test between forget and retain:

	Statistic	Captures
$\mathcal{M}_\mu$	$\bar{x}_i = \frac{1}{\lvert\mathcal{S}\rvert}\sum a_i^{(s)}$	Mean activation
$\mathcal{M}_\sigma$	$\sqrt{\mathrm{Var}(a_i^{(s)})}$	Volatility
$\mathcal{M}_{\mu/\sigma}$	$\bar{x}_i/(\sigma_i+\epsilon)$	Signal-to-noise
$\mathcal{M}_\gamma$	$\mathbb{E}[(a_i-\bar{x}_i)^3]/\sigma_i^3$	Skewness
$\mathcal{M}_{\mathrm{KL}}$	$\frac{1}{2}(D_{\mathrm{KL}}(p_f\Vert p_r)+D_{\mathrm{KL}}(p_r\Vert p_f))$	Distribution

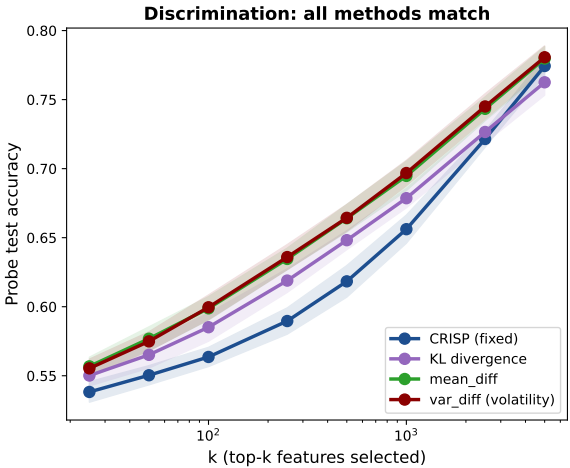
Each moment asks a different question:

- Does the feature fire **more** for forget users?
- Does it fire more **variably**?
- Does it fire more **consistently**?
- Does it fire with different **asymmetry**?
- Are the full **distributions** different?

Top-k features by each score are selected for targeted suppression during LoRA fine-tuning.

Different features distinguish users in different ways. No single statistic captures all of them.

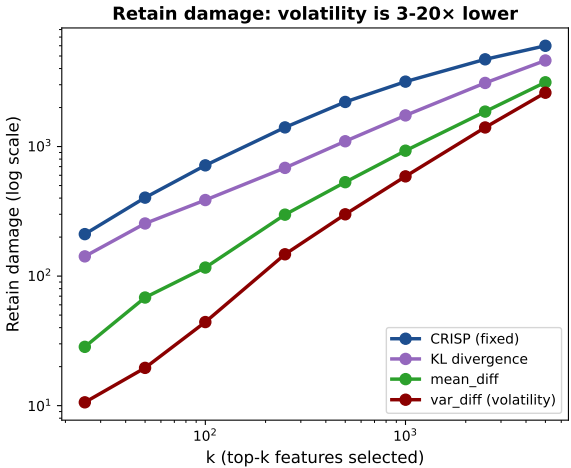
All Moment Selectors Match CRISP Discrimination



Can moment-selected features identify forget users?

- A logistic probe trained on the top-k selected features predicts forget vs retain membership
- All moment selectors reach comparable accuracy to CRISP
- Discrimination improves with more features, saturating around $k = 2500$

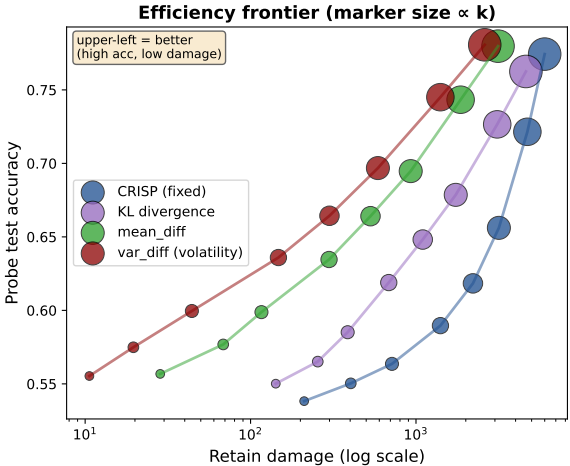
Moment Selectors Cause 3–20× Less Retain Damage



How much collateral damage do the selected features cause?

- Retain damage = total activation mass of selected features on retain users
- Volatility-based selection ($\mathcal{M}_\sigma$) causes **3–20× less damage** than CRISP at every k
- Mean-based selection also substantially lower

Moment Selectors Dominate the Efficiency Frontier

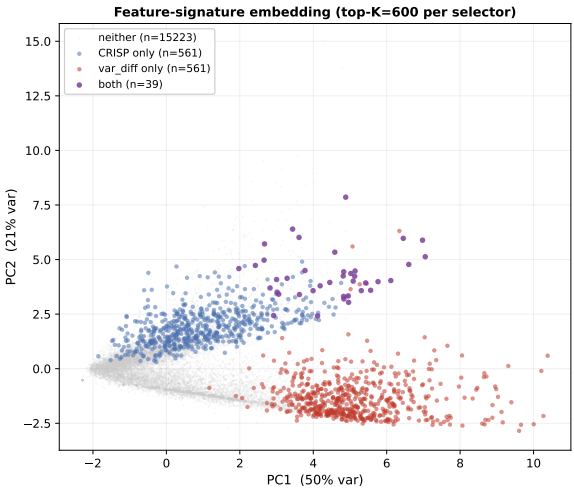


Which selector gives the best accuracy per unit of damage?

- Upper-left is ideal: high discrimination, low retain damage
- Moment selectors achieve the **same accuracy as CRISP** while sitting further left (less damage)
- Volatility ($\mathcal{M}_\sigma$) is consistently the most efficient selector

Moment selectors find features that are equally diagnostic but less entangled with retain users.

CRISP and Moment Selectors Find Different Features

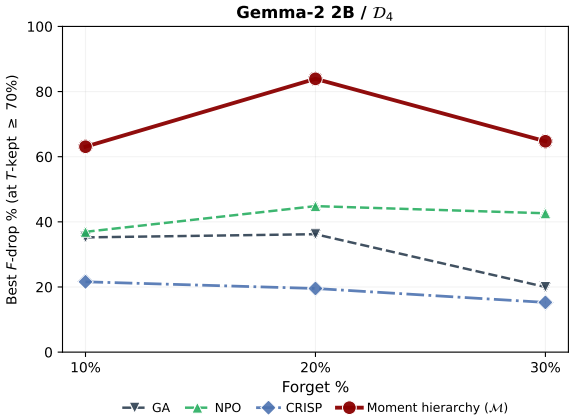


Do different selectors find the same features?

- Each dot = one SAE feature, projected by its statistical fingerprint
- CRISP (blue) and $\mathcal{M}_\sigma$ (red) cluster in **different regions**
- Only **39 out of 600** features overlap (3.4% Jaccard)

Different selectors surface complementary signal. NOT just different rankings of the same features.

Our Method is Consistent Across Deletion Sizes

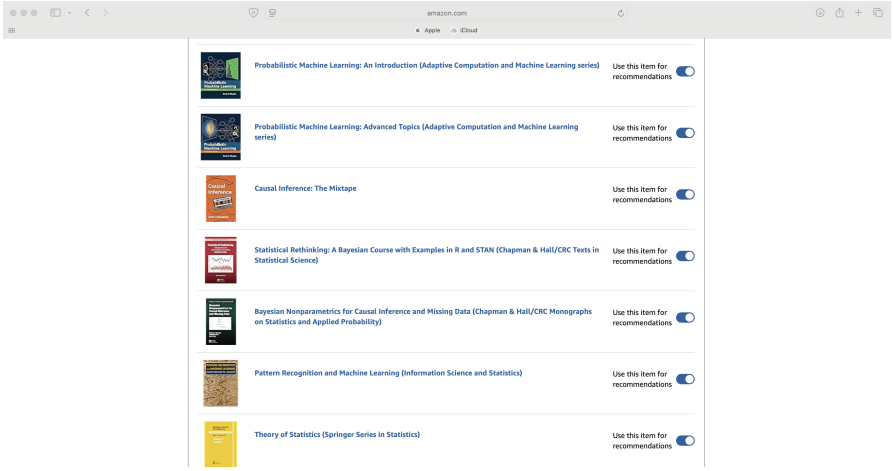


Does performance hold as the deletion set grows?

- Moment hierarchy maintains **60–84%** forgetting at 10%, 20%, and 30% deletion
- Baselines plateau or degrade at larger deletion sizes

Appendix

Amazon: “Improve Your Recommendations”



Back

Profit Score Framework Details

Utility Retention Metrics (Revenue Protection):

- NDCG@K, Recall@K, Hit@K
- Measures how well the model still recommends after unlearning

Computational & Operational Cost:

- Retraining time / GPU hours
- Latency per deletion request
- Energy cost for large-scale deletions

Leakage Risk:

- Membership Inference Attack (MIA) success rate
- Post-unlearning: should approach random chance ($\sim$50% AUC)
- Unlearning Accuracy (UA): Drop in accuracy on forget set

Back

ComScore Data Details

Data Tables:

- `comscore_search_fact`: Search phrases
- `comscore_url_traffic`: Domains visited
- `comscore_category_map`: Site categories
- `purchase_items`: Transaction records

Top Categories:

- Home & Living: 438K events
- Electronics & Computing: 20K events
- Apparel & Accessories: 17K events
- Books, Music & Video: 11K events

User Behavior:

- Avg actions per user: 14.12
- Most active user: 2,170 actions
- Avg basket size: 2.98 items
- Avg basket value: $3,317

Seasonality:

- Peak: Month 1 (78K events)
- Trough: Month 9 (39K events)
- Holiday uptick: Months 11–12

Back

Why ComScore is Uniquely Suited for Unlearning Research

Scale

111M searches
40K items
Multi-platform

Full Pipeline

Search $\rightarrow$ Click $\rightarrow$
Browse $\rightarrow$ Purchase
Causal chain preserved

Temporal

Per-second granularity
Session reconstruction
Query reformulations

Key Insight for Unlearning

User influence is **not instantaneous**—it accumulates over time. Removing a user requires removing a **temporally ordered chain of influence**, not a single data point.

Back

How is Unlearning Done? SISA & RecEraser

SISA (Exact Unlearning):

- Shard data, train sub-models
- Retrain only affected shard
- Problem: Breaks collaborative signals

RecEraser (For RecSys):

- Balanced partitioning (not random)
- Preserves user-item clustering
- 10–27× faster than full retrain

SISA

Random sharding → Breaks collaborative graph

Majority voting → Accuracy degradation

RecEraser

Balanced partitioning → Preserves local signals

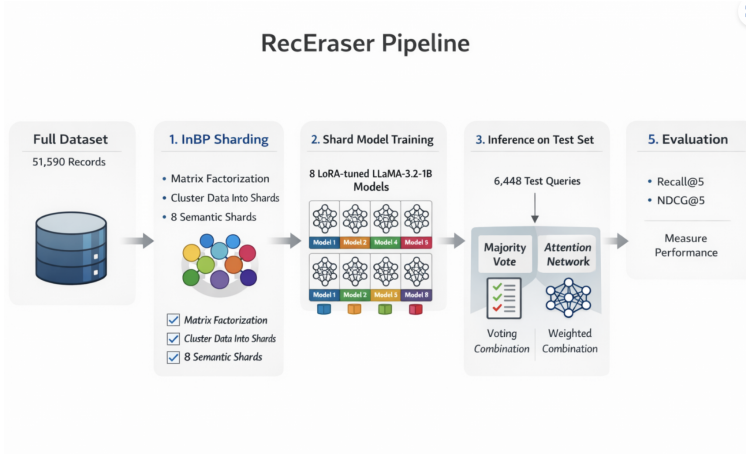
Attention aggregation → Maintains accuracy

Key Limitation

Both assume user influence can be isolated, which does not hold in dense, sequential recommender data.

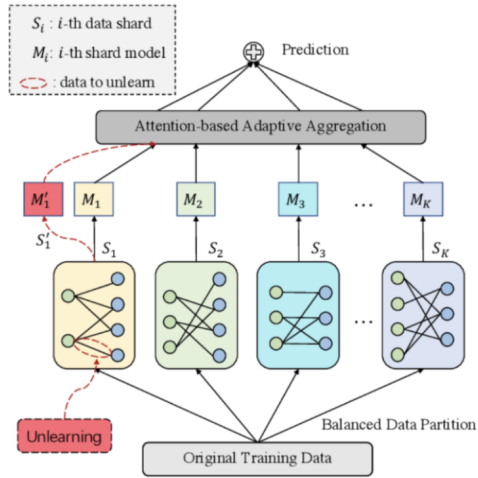
Back

RecEraser Pipeline



Back

SISA and RecEraser Architecture



Back

Key Related Work

Paper	Year	Method
SISA (Bourtoule et al.)	2021	Sharded training
RecEraser (Chen et al.)	2022	Balanced partitioning for RecSys
UltraRE	2023	Error decomposition
CURE4Rec	2024	Benchmark for RecSys unlearning
CRISP	2025	SAE-based concept removal
SAE Subspace Projections	2025	SAE-guided parameter updates

SAE Unlearning Literature:

- "Applying Sparse Autoencoders to Unlearn Knowledge in Language Models" (2024)
- "Sparse-Autoencoder-Guided Internal Representation Unlearning for LLMs"
- "SAEs Can Improve Unlearning: Dynamic SAE Guardrails" (2025)

Back

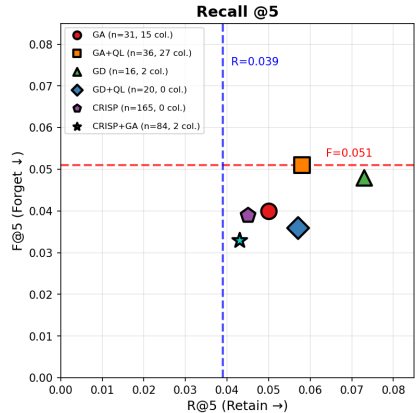
Comparison with Approximate Unlearning Methods

Retain & Forget Performance

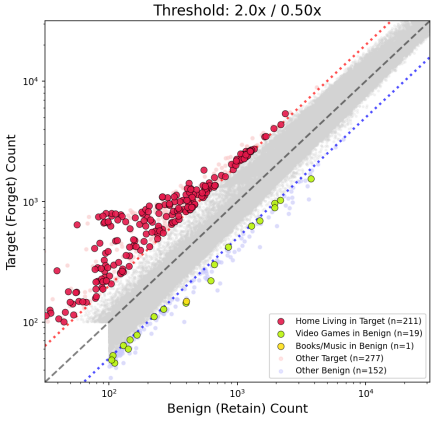
Method	@5 R	@5 F	@10 R	@10 F
GA	.050	.040	.055	.044
GA+QL	.058	.051	.062	.058
GD	.073	.048	.082	.054
GD+QL	.057	.036	.065	.042
CRISP	.045	.039	.053	.042
CRISP+GA	.043	**.033**	.050	**.037**
Baseline	.039	.051	.045	.057

Key Finding

CRISP+GA: Best forget rate (**-36.7%**) while retaining model utility (+12.0%)



Feature Selection via Sparse Autoencoders



Features colored by product category

Why SAEs for Unlearning?

- Features map to **semantic concepts**
- Salient features cluster by category
- Enables **targeted suppression**

Interpretability Advantage

Gradient-only: which weights changed?
SAE-guided: which concepts suppressed?

Back

Detailed Baseline Comparison: RecEraser vs SISA

Condition	RecEraser		SISA	
	Recall@5	NDCG@5	Recall@5	NDCG@5
Baseline	5.71	4.05	3.59	2.98
10% unlearned	3.89	2.20	1.93	1.62
20% unlearned	4.03	2.26	2.40	1.79
30% unlearned	3.80	2.15	—	—

RecEraser Advantage

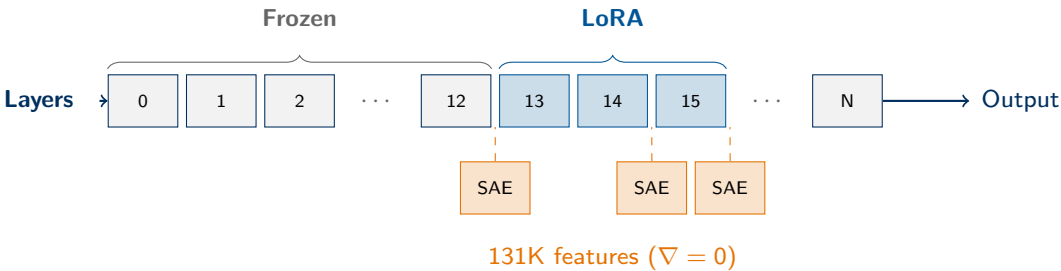
- Higher baseline performance
- Better retention after unlearning
- Balanced partitioning helps

SISA Weakness

- Random sharding hurts RecSys
- Larger performance drop
- Doesn't scale to LLMs well

Back

Framework: Layer Architecture



Frozen Layers

Preserve general knowledge

LoRA Layers

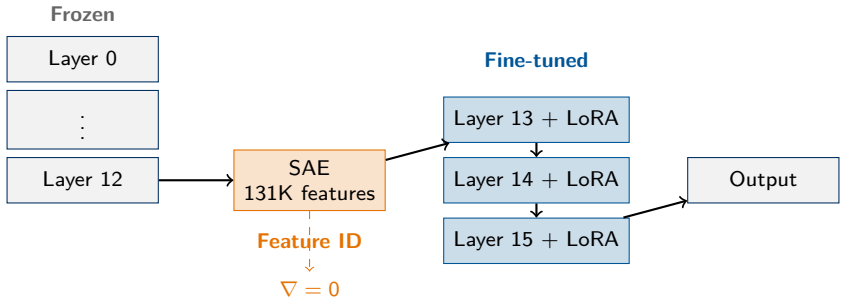
Learn to suppress user-specific features

SAEs

Identify which features to target

Back

Framework: Data Flow



Frozen Layers

Layers 0–12: Preserve general knowledge

SAE

Identifies user-specific salient features

LoRA

Suppresses salient activations

Back