جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

NUS National University of Singapore

JOHNS HOPKINS UNIVERSITY

Towards Lifecycle Unlearning Commitment Management:

Measuring Sample-level Unlearning Completeness

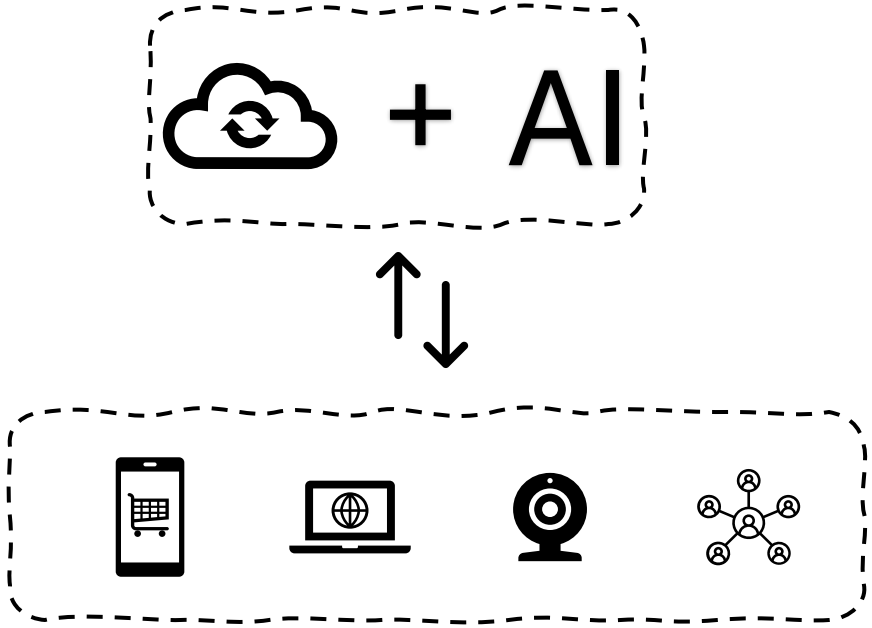
Cheng-Long Wang[†], Qi Li[†, ‡], Zihang Xiang[†], Yinzhi Cao[§], Di Wang*[†]

[†] King Abdullah University of Science and Technology

[‡] National University of Singapore [§] Johns Hopkins University

What’s Machine Unlearning

'Machine Unlearning' is the reverse process of machine learning. It aims to modify a trained model to ‘forget’ specific data, so that the updated model behaves as if it had never been trained on the forgotten data.

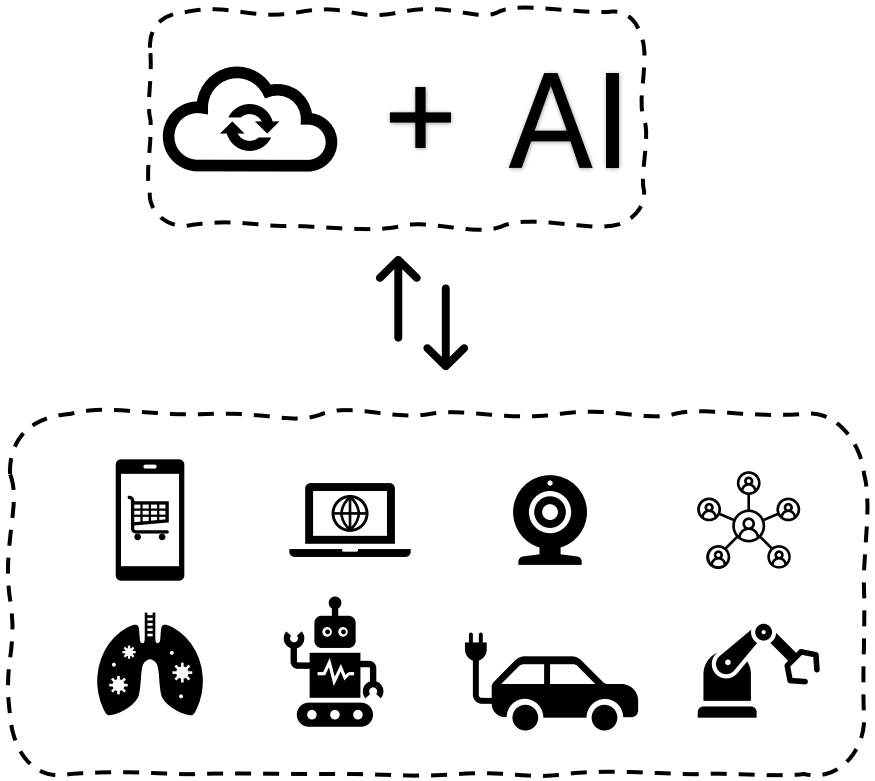


Why Machine Unlearning Matters

- Server side: data & model management.
  - forget harmful or low-quality data (erotic, biased, noisy)
  - Unlearn copyrighted or policy-sensitive content (alignment & compliance)
  - Data market: licensing, rental, and revocable sharing
  - Life-cycle: Remove outdated knowledge as environments or policies change

- User side: privacy & consent
  - Remove Personally Identifiable Information (PII)
  - Change of user consent on data usage (Right to be forgotten)

- System side: security & reliability
  - Remove backdoors, poisoned updates, or leaked secrets
  - Reduce hallucinations and other unreliable outputs



Implementing Machine Unlearning

- ❑ Exact Machine Unlearning
  - ❑ Horizontal: SISA, Subsampling
  - ❑ Vertical: reload saved checkpoints
- ❑ Approximate Machine/LLM Unlearning
  - ❑ Parameter-side: influence function, gradient ascent, knowledge editing, fisher noise…
  - ❑ Loss-side: fine-tuning with redesigned loss terms (forget loss, retain loss, …)
  - ❑ Activation-side (Inference time): embedding corruption, neural activation redirection
  - ❑ Prompt-side (Inference time): in-context unlearning, pretending not to know, filtering …
- ❑ Unlearning Measurements/Auditing

Security by Design

Low cost/Post-hoc

Verifiability/Accountability

Unlearning Methods: Exact vs. Approximate

Machine Unlearning: removing data influence from the model without costly full retraining

- **Exact unlearning**: redesign the training pipeline and partially retrain.
- **Approximate unlearning**: adjust parameters/outputs to mimic the retrained model.

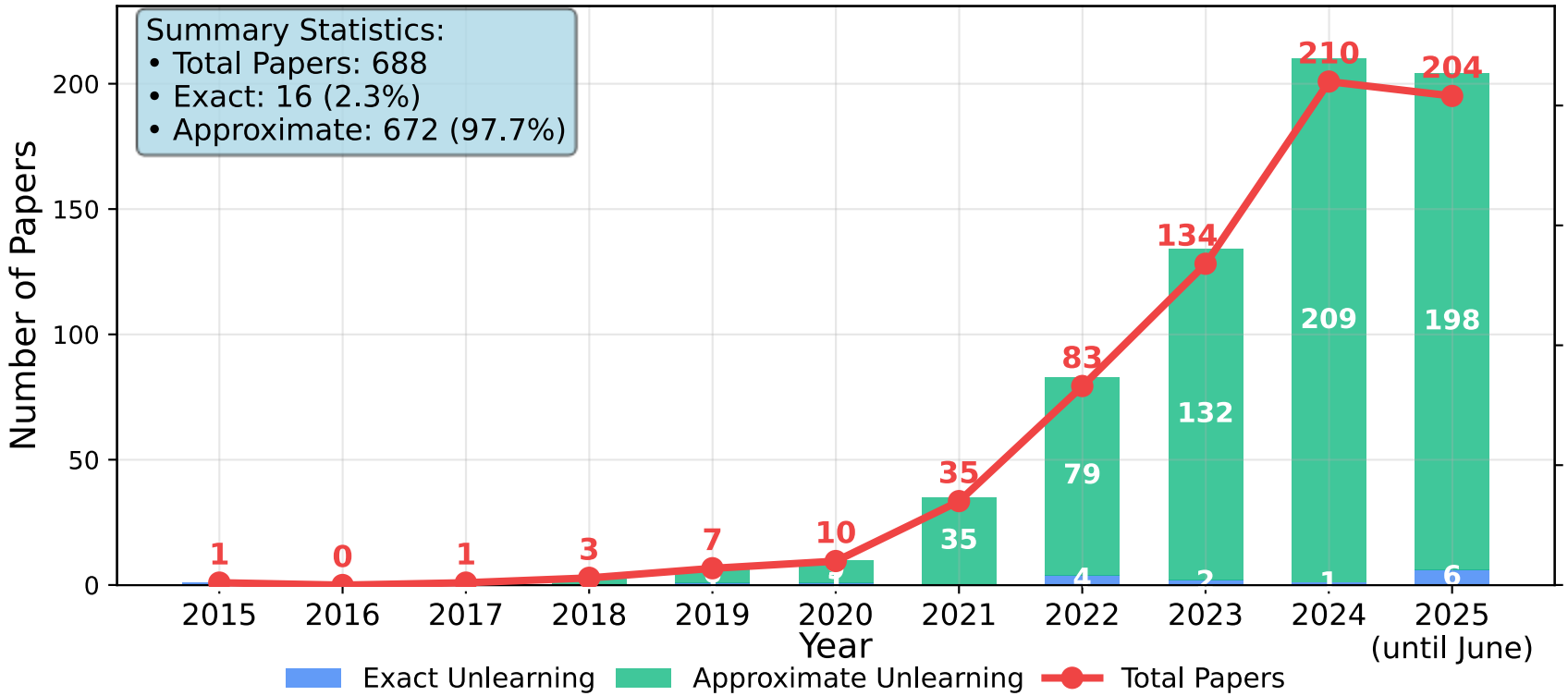


Figure 1: Unlearning papers over the past 10 years

Unlearning Methods: Exact vs. Approximate

- **Machine Unlearning:** removing data influence from the model without costly full retraining
  - **Exact unlearning**: redesign the training pipeline and partially retrain.
  - **Approximate unlearning**: adjust parameters to mimic the retrained model.

When they claim to forget,
how well do they ACTUALLY forget?

Figure 1: Unlearning papers over the past 10 years

Revisiting How Outsiders Tell Membership

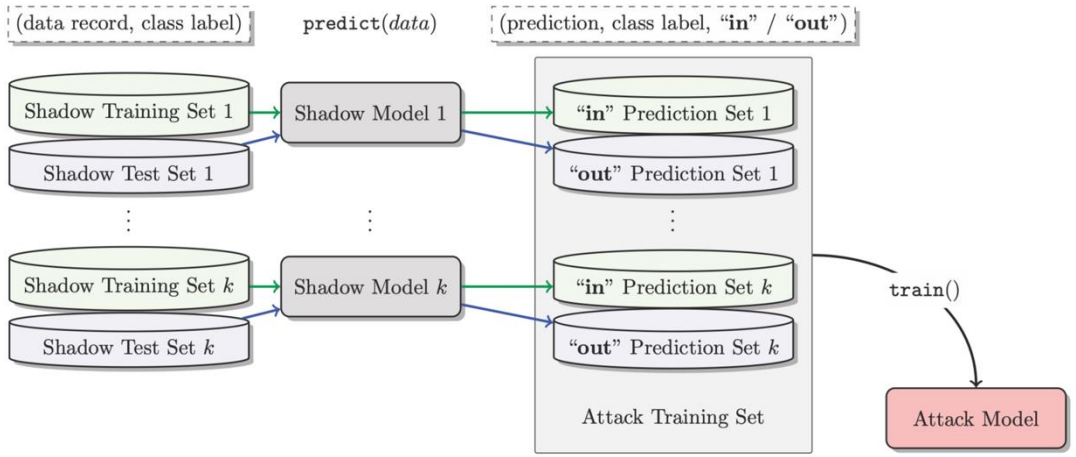


Figure 2: Training the Membership Inference Attack (MIA) attack model on the inputs and outputs of the shadow models[1] .

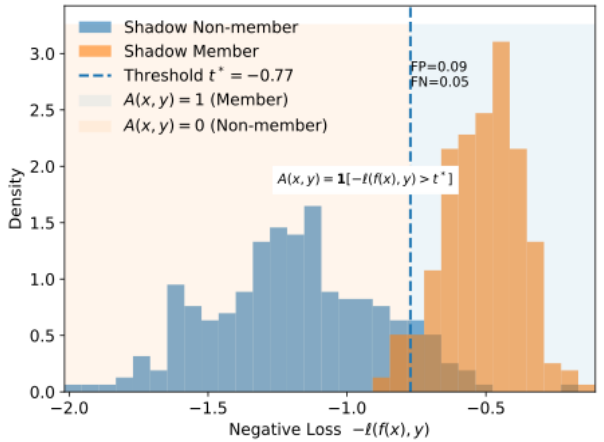


Figure 3: Loss-based MIA using one threshold for all queries (Offline Attack).

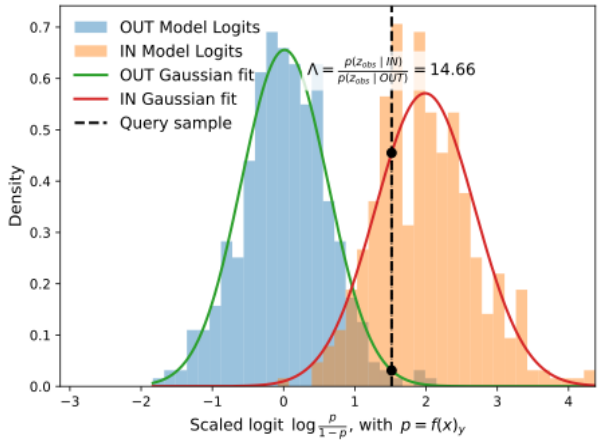
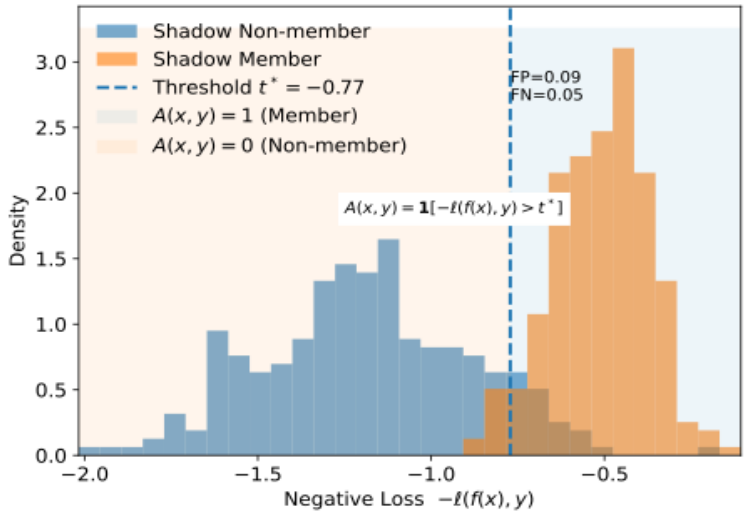


Figure 4: Likelihood-ratio test attack: train IN/OUT models for each query (Online Attack).

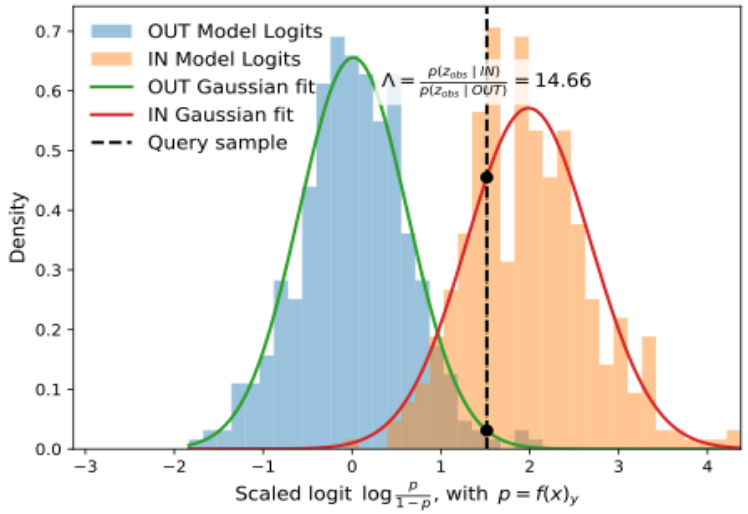
[1]R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership Inference Attacks Against Machine Learning Models, IEEE S&P 2017.

MIA Flaws for Unlearning Inference

Why MIAs Are Not Suitable for Unlearning Inference



Offline MIA often flags generalized-well samples as 'not unlearned'
→ false under-unlearning signals (low TNR)



Online MIA cost grows with each query
→ infeasible for large-scale unlearning

Our First Step — Better Modeling of OUT Responses

❑ **A uniform-scale transformation: GumbelMap**: $r(z;\theta) = -\log(-\log(p))$

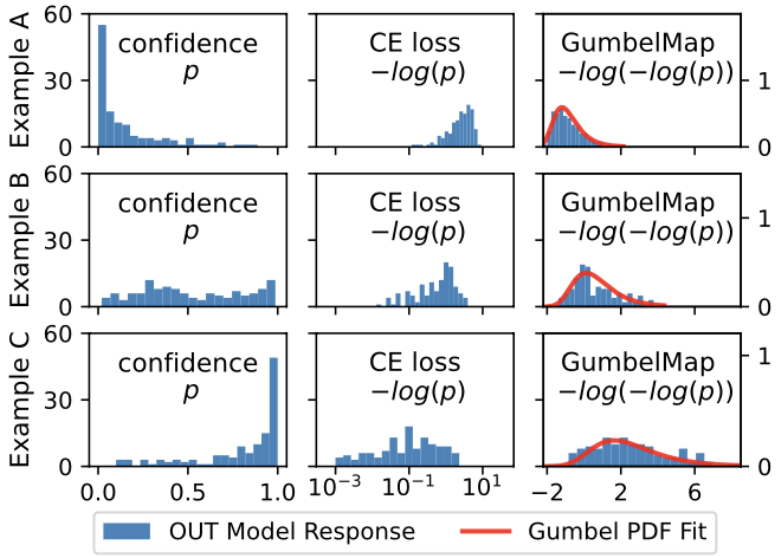


Figure 3: Histograms of OUT model responses. We trained 128 models on random subsets of CIFAR-100 and plotted the model responses for three examples that were not part of the training data for these models.

- Stable scale, unlike CE loss spanning orders of magnitude.
- Well fit by Gumbel distribution, enabling efficient parametric modeling the per-example 'easy-to-generalize' scores
- More sensitive as model confidence → 1, better differentiating high-confidence samples

$$\alpha = \mu - \gamma \cdot \beta,$$

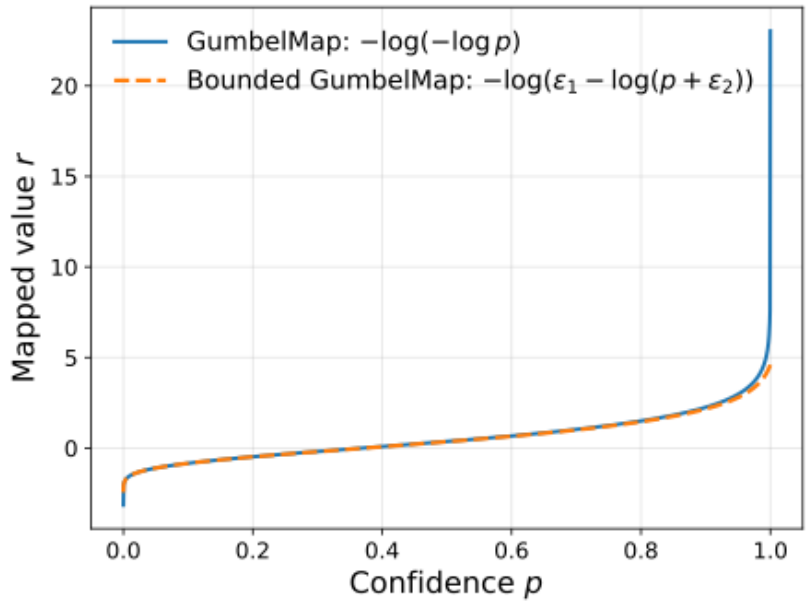
$$\beta = \sqrt{\frac{6\sigma^2}{\pi^2}},$$

$$f(x;\alpha,\beta) = \frac{1}{\beta} e^{-\frac{x-\alpha}{\beta}} e^{-e^{-\frac{x-\alpha}{\beta}}},$$

$$F(x;\alpha,\beta) = e^{-e^{-\frac{x-\alpha}{\beta}}}.$$

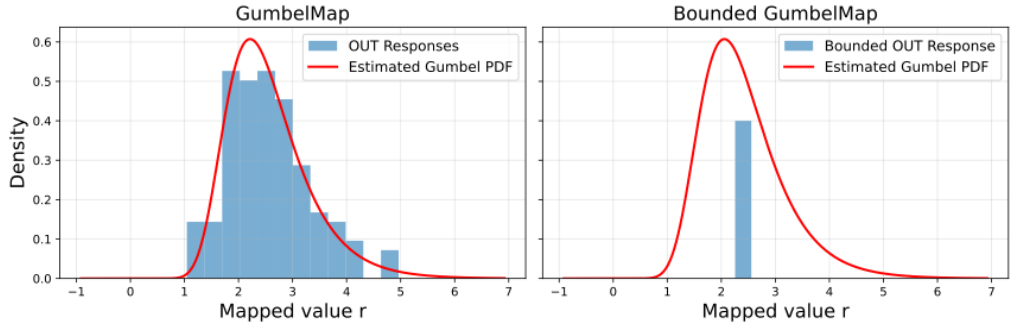
Bounded GumbelMap

- A more reliable transformation: Bounded GumbelMap: $r(z;\theta) = -\log(\varepsilon_1 - \log(p + \varepsilon_2))$



Transforming model confidence with GumbelMap and its bounded variant.

- GumbelMap becomes overly sensitive as p→ 1, while the bounded variant remains controllable.
- Bounded GumbelMap yields bounded outputs, and with Popoviciu's inequality we can estimate parameters reliably even from a single shadow model.



Variance trick: cross-example variance ≈ cross-model variance.

A New Question: What if Forgetting Isn't Binary?

Exact Unlearning → Binary inclusion → Binary Unlearning Inference

- $b_i = 1$: Model response to i-th sample reflects memorization.
- $b_i = 0$: Model response to i-th sample rely entirely on generalization.

Approximate Unlearning → from binary to spectrum → Score-based Unlearning Inference

- $s_i \in [0,1]$: model response lies between on a spectrum between generalization and memorization.

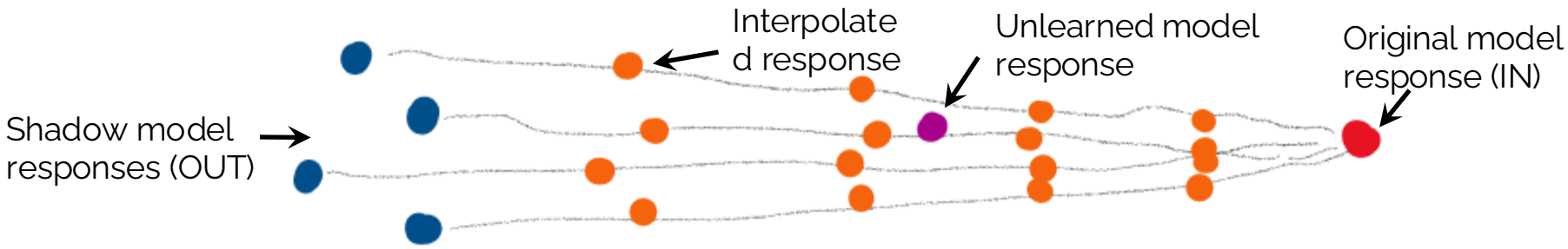
$$\hat{b} = \mathbb{1}[Score(z;\theta) > \tau]$$

Binary Membership Status

$$Score(z;\theta)$$

Score-based Membership Status

Our Second Innovation: Response Interpolation



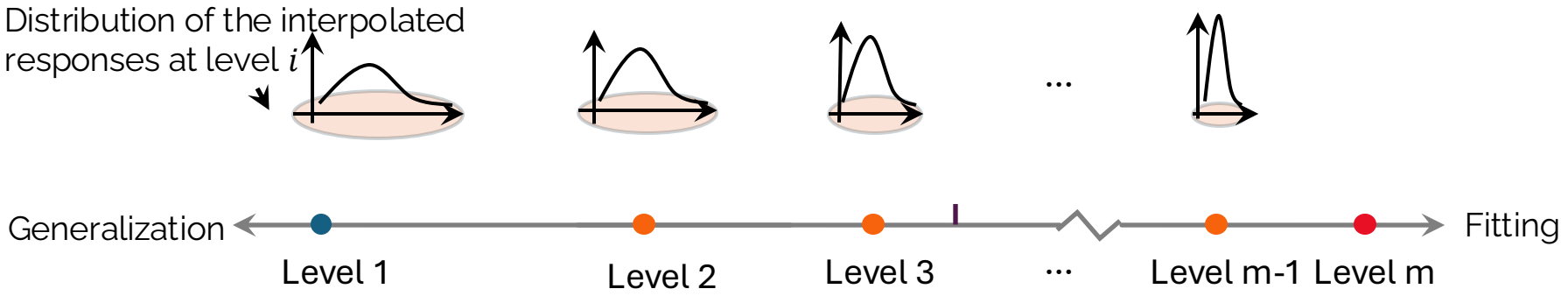
Core Equation: Suppose $\boldsymbol{r(z;\theta)}$ is model $\boldsymbol{\theta}$'s response to sample $\boldsymbol{z}$, we compute the $\boldsymbol{i}$-th interpolation $\boldsymbol{r_i(z;\tilde{\theta},\theta)}$ between two models as:

$$\boldsymbol{r_i = \frac{m-i}{m-1} r(z;\tilde{\theta}) + \frac{i-1}{m-1} r(z;\theta), \text{where } i \in \{1,\dots,m\}}$$

What it does:

- Connects shadow model response $\boldsymbol{r(z;\tilde{\theta})}$ to original model response $\boldsymbol{r(z;\theta)}$
- Creates $\boldsymbol{m}$ behavior points per trajectory (shadow-original model pair)
- No online training

From Interpolation to Membership Score



⌨ **For each level i :** Assuming responses $\boldsymbol{r_i}$ at level $\boldsymbol{i}$ followa Gumbel distribution $\boldsymbol{\Phi_i}$, we estimate the probability of unlearned model's response $\boldsymbol{r' = r(z; \theta')}$ larger than responses from $\boldsymbol{\Phi_i}$:

$$\boldsymbol{q_i = \Pr[r' > X], \text{where } X \sim \Phi_i}$$

📦 **Weighted average:** Discount noisy early levels (close to pure generalization)

$$Score(z; \boldsymbol{\theta}, \boldsymbol{\theta'}) = \frac{\sum_{i=1}^{m-1} i \cdot q_i}{\sum_{i=1}^{m-1} i}$$

Interpolated Approximate Measurement (IAM)

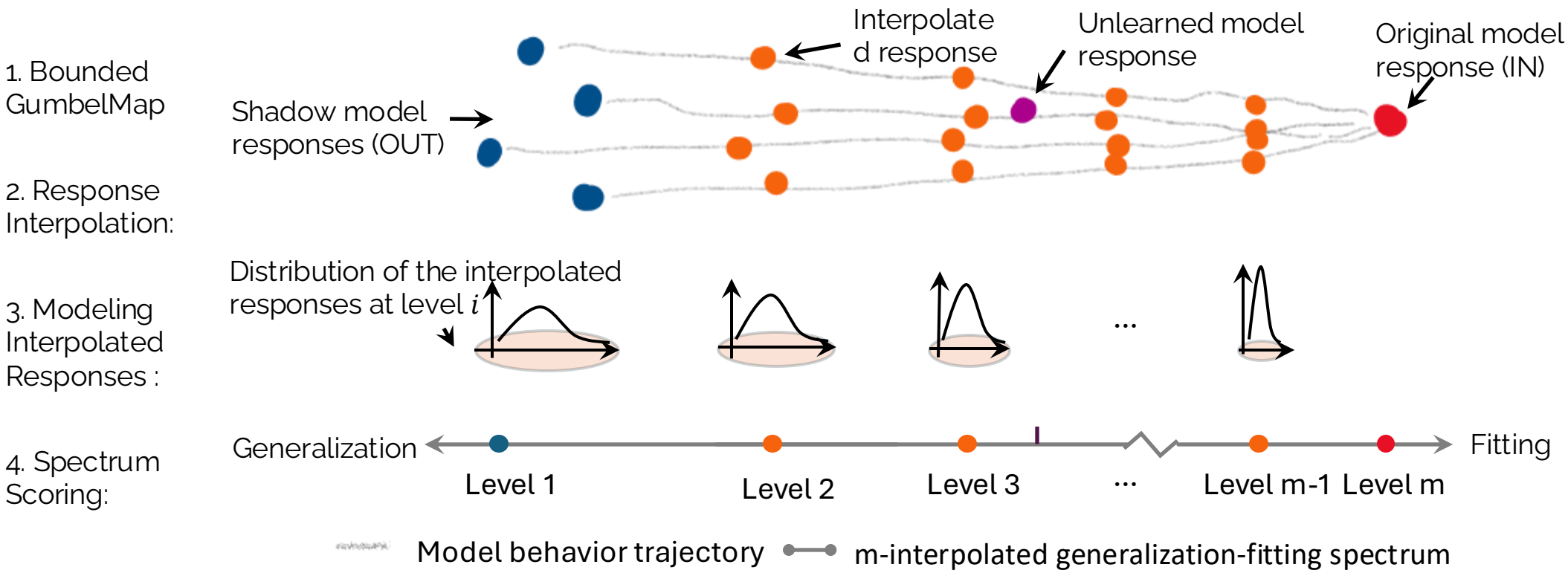


Figure 2: Interpolated Approximate Measurement Framework

Based on the IAM framework, we designed a bounded signal mapping to estimate unlearning completeness with one reference model and theoretical support.

Binary/Score-Based Unlearning Inference Performance

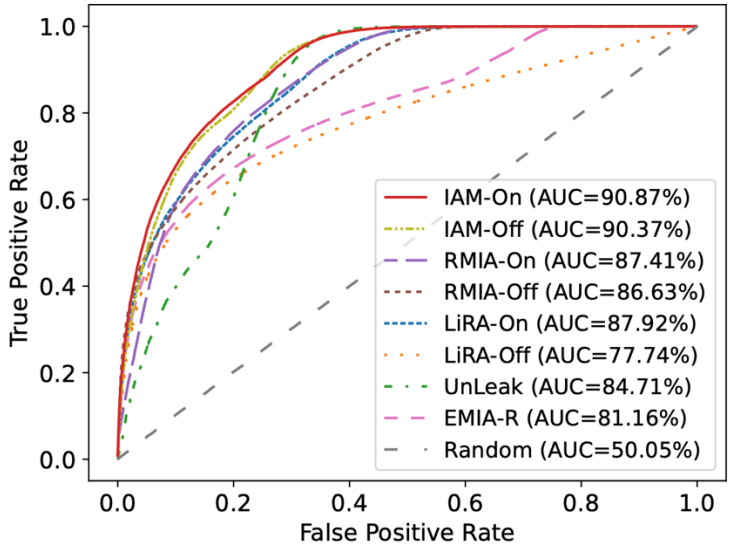


Figure 1: ROC curve of IAM versus prior MIA methods (RMIA [63], EMIA [62], LiRA [6], Unleak [10]) for exact unlearning inference on CIFAR-100. Suffixes -ON and -OFF denote online and offline variants of methods, respectively. Evaluation involved randomly unlearning 500-sample batches (average of 10 runs). All methods are limited to **one pre-trained shadow model** (used as the OUT model). For online variants, the original model serves as the IN model.

Binary Unlearning Inference Results

Table 2: Spearman correlation of all methods on ScoreUI tasks. See Appendix D.6 of the supplementary material 7 for Purchase dataset results.

	Method	CIFAR-10	CIFAR-100	CINIC-10
Offline	Random	0.000±0.000	-0.000±0.000	0.000±0.000
	EMIA-P	0.053±0.000	0.440±0.000	0.141±0.000
	EMIA-R	0.347±0.001	0.358±0.003	0.500±0.001
	LiRA-Off	-0.177±0.013	0.341±0.019	-0.021±0.022
	RMIA-Off	-0.339±0.002	0.408±0.002	-0.020±0.004
	IAM-Off	**0.480±0.002**	**0.713±0.001**	**0.649±0.001**
Online	UpdateAtk	0.268±0.003	0.430±0.007	0.336±0.003
	UnLeak	0.430±0.003	0.672±0.005	0.559±0.127
	LiRA-On	0.247±0.003	0.637±0.011	0.452±0.006
	RMIA-On	-0.359±0.001	0.405±0.001	-0.075±0.002
	IAM-On	**0.480±0.002**	**0.713±0.001**	**0.647±0.001**

Score-based Unlearning Inference Results

Example of Spectrum Inference

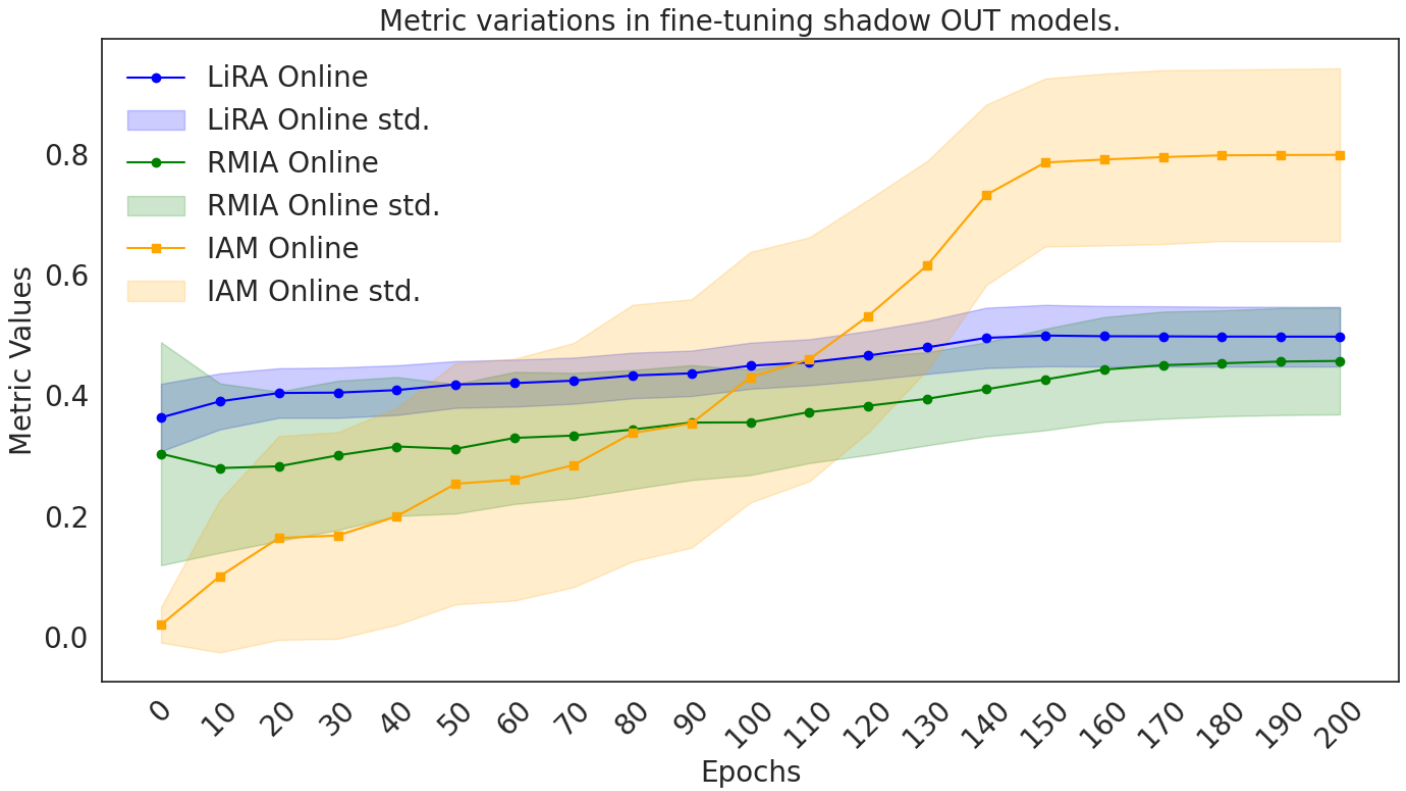


Figure 6: Example of Spectrum Inference. We simulate the transition from generalization to fitting by continuing to train shadow OUT models on the original set until they reach the target model's high training accuracy.

LLM Results

Table 6: Unlearning inference results on Llama-2 7B model.

Method	Results	
	AUC(%)	Spearman
Bag-of-Words	54.11 ± 0.81	0.037 ± 0.004
Loss	87.46 ± 0.52	0.385 ± 0.003
Zlib	89.28 ± 0.46	0.407 ± 0.003
Ratio	91.26 ± 0.45	0.458 ± 0.003
SURP	87.19 ± 0.54	0.336 ± 0.003
Min-K% Prob	87.84 ± 0.53	0.167 ± 0.004
Min-K%++	84.69 ± 00.57	0.336 ± 0.003
LiRA-On	92.99 ± 0.37	0.425 ± 0.003
RMIA-On	92.78 ± 0.34	0.508 ± 0.003
IAM-On	**93.55 ± 0.33**	**0.509 ± 0.003**

Adopting IAM on LLM

Table 7: Predicted membership scores of ScoreUI on three unlearned Llama-2 7B models (exact unlearning, negative-label unlearning, and refusal-prefix unlearning).

$\mathbf{b}_i$	Method	Retrain	Negative	Refusal
1	LiRA-On	0.25 ± 0.04	0.06 ± 0.03	0.07 ± 0.03
	RMIA-On	0.44 ± 0.40	0.24 ± 0.12	0.24 ± 0.12
	IAM-On	0.84 ± 0.22	0.88 ± 0.14	0.88 ± 0.13
0	LiRA-On	0.18 ± 0.03	0.07 ± 0.03	0.08 ± 0.03
	RMIA-On	0.02 ± 0.04	0.26 ± 0.13	0.24 ± 0.14
	IAM-On	0.33 ± 0.18	0.91 ± 0.12	0.90 ± 0.11

* Red: Under-unlearning risks.

Adversarial Bypass Scenario

Benchmarking Approximate Unlearning Methods

Table 8: Approximate Unlearning Results on CIFAR-100 and CINIC-10; additional results for CIFAR-10 and Purchase datasets are provided in Appendix D.7 of the supplementary material 7.

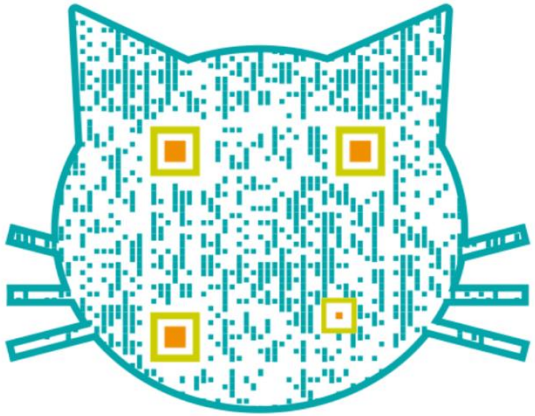
Method	CIFAR-100		CINIC-10	
	$\mathbf{b}_i = 1$	$\mathbf{b}_i = 0$	$\mathbf{b}_i = 1$	$\mathbf{b}_i = 0$
Retrain	0.81±0.16	0.01±0.01	0.72±0.17	0.00±0.01
Fine-tune	0.79±0.17	0.32±0.25	0.65±0.21	0.01±0.02
Ascent	0.80±0.16	0.59±0.23	0.73±0.17	0.68±0.19
L-codec	0.71±0.33	0.41±0.31	0.67±0.22	0.04±0.10
Boundary	0.83±0.15	0.50±0.26	0.65±0.24	0.03±0.09
Forsaken	0.82±0.15	0.57±0.24	0.73±0.17	0.66±0.20
SSD	0.79±0.19	0.01±0.01	0.65±0.24	0.03±0.09
Fisher	0.83±0.15	0.01±0.01	0.65±0.24	0.01±0.01

* Red: Under-unlearning, Orange: Over-unlearning.

Threshold-based Unlearning Risk Identification

Conclusion

- **We define the task of unlearning inference and propose Interpolated Approximate Measurement (IAM) – a framework natively designed for this task.**
- **We prove that IAM enables reliable scoring with just one shadow model, using variance bounds and Popoviciu's inequality to justify its estimation efficiency.**
- **We benchmark approximate unlearning methods and scale IAM to LLM, showing that unlearning risks are measurable and pervasive.**



https://github.com/Happy2Git/Unlearning_Inference_IAM

Thanks~