In-Context Probing for Membership Inference in Fine-tuned Language Models

Zhexi Lu, Hongliang Chi, Nathalie Baracaldo, Swanand Ravindra Kadhe,
Yuseok Jeon, Lei Yu

Rensselaer Polytechnic Institute
IBM Research
Korea University

Presenter: Zhexi Lu

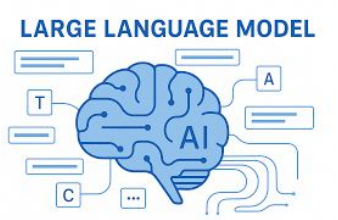
Paper Link:

LLMs are fine-tuned to fit various domains

AI Scientist

Scientific data

Finance data



LLM Backbone

Medical data

Legal case data

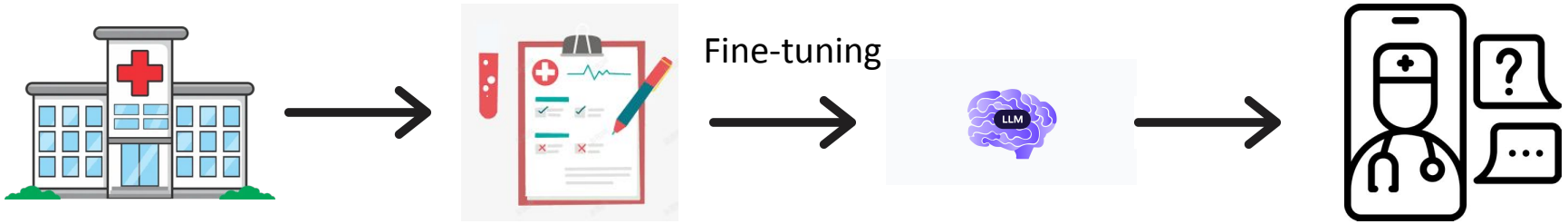
AI Doctor

AI in Finance

Pre-trained LLM + **Domain Data (often privacy-sensitive)** →**Fine-tuning** → Domain-specific LLM

AI Lawyer

Vulnerability in Fine-tuned LLM

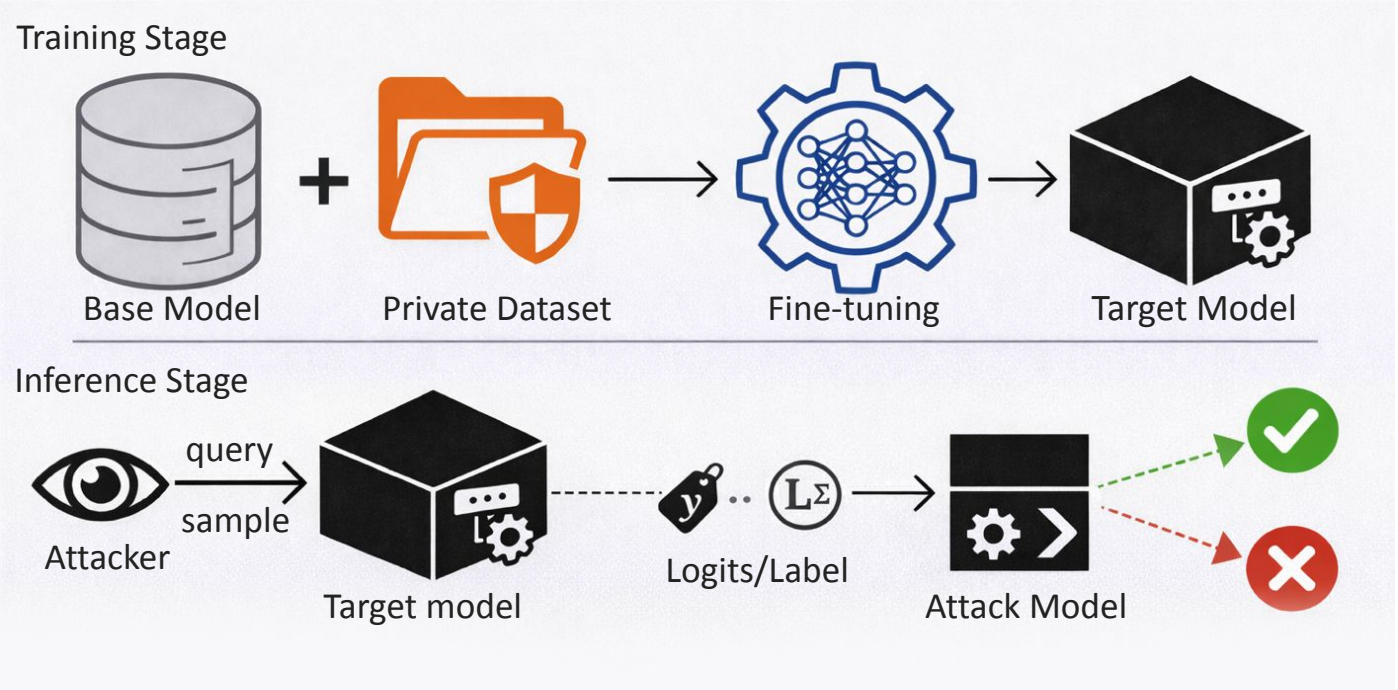


Could the training data leak private information through this interface?

Data Extraction Attacks: prompts the model to regurgitate verbatim training samples (e.g., patient names, diagnoses)

Membership inference Attacks: infers whether a specific individual's data was included in the training set.

MIA against Fine-Tuned LLM



Existing MIA on LLM

Reference-Free Methods:

- Min-K%: Extract the log-probabilities of the lowest k% of tokens as a membership signal.
- Neighborhood Attack: Compare the difference between log-probability of the target sample and the average of its neighborhoods as a membership signal.
- ReCaLL: prefixing target samples with non-member context causes a greater reduction in log-likelihood for member data than for non-member data.

Reference-based Methods:

- SPV-MIA: Calibrate difficulty by self-prompting the model to generate reference data, which is then used to train a reference model.

Limitations:

(1) Rely on heuristic post-training signals (token confidence, log-likelihood etc), which can be brittle and lack principled grounding.

(2) Some methods require training reference models, which is computationally expensive for LLMs.

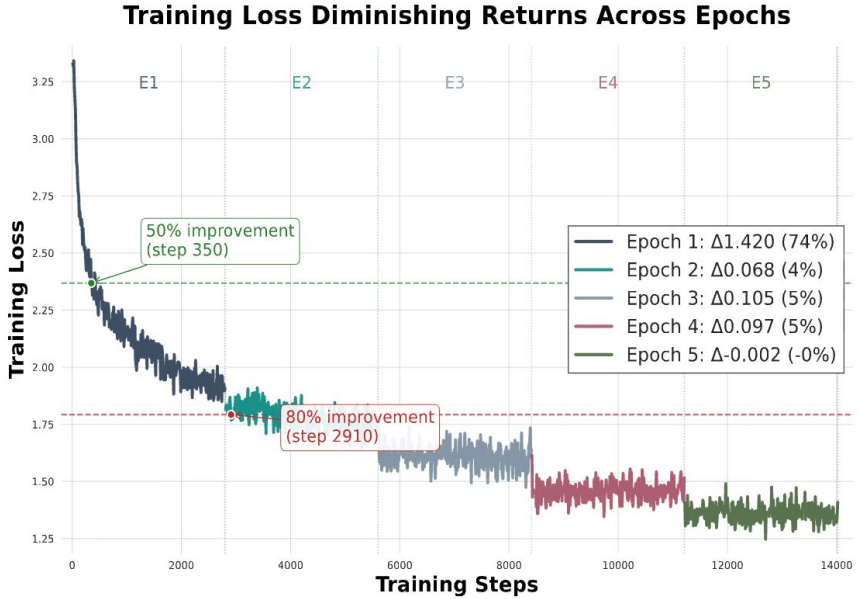
Our Approach

- Reference free
- Exploiting fundamental signal from training dynamics

Design Principles

- As training progresses, the decrease in loss exhibits a diminishing return effect.

What are the differences in behavior between non-members and members in training stage?



Design Principles

Empirical Study

1. We fine-tuned the base model LLama3.2-3B-instruct for two epochs on the HealthCareMagic dataset and then randomly selected 1,000 member samples and 1,000 non-member samples from the test dataset.

2. We combined these samples, SFT with same settings, and measured the per-sample loss reduction.

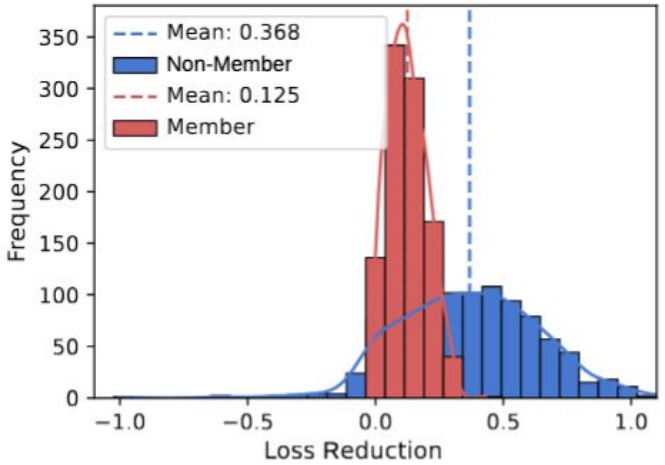


Fig. 3: Fine-tuning with Members V.S. Non-Members

Design Principles

Empirical Study

- The loss reduction for member data mainly between 0 and 0.25, while that for non-member data decreased mainly between 0 and 1.

We define this phenomenon as the "Optimization Gap" which can be used for Membership Signal.

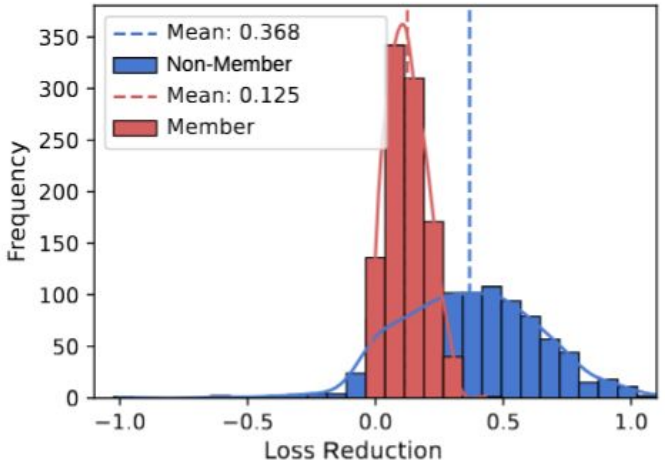


Fig. 3: Fine-tuning with Members V.S. Non-Members

Design Principles

Empirical Study

- The loss reduction for member data mainly between 0 and 0.25, while that for non-member data decreased mainly between 0 and 1.

We define this phenomenon as the "Optimization Gap" which can be used for Membership Signal.

We can't fine-tune the Target model！
How can we get this in Black-Box Setting?

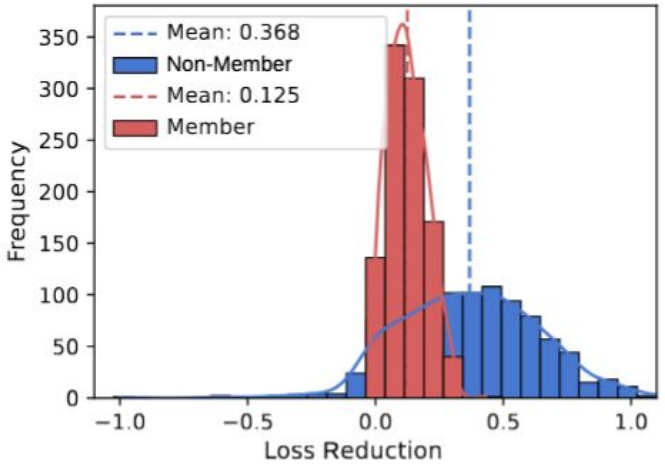


Fig. 3: Fine-tuning with Members V.S. Non-Members

From True Optimization to In-Context Approximation

- *ICL as a form of implicit optimization*, where the model internally simulates gradient-based adjustments in response to the provided context [1].
  - This view is further supported by data attribution studies[2], which show that carefully designed context perturbations can approximate gradient-based influence scores.

Can we leverage the model's in-context learning capabilities to simulate the behavior after fine-tuning?

$$\Delta_{\mathrm{LL}}(s) = LL(y \mid x; \mathcal{M}) - LL(y \mid x; \mathcal{M}') \quad ? \quad \mathrm{ICP}_{\mathrm{score}}(s, C) = LL(y \mid x; \mathcal{M}) - LL(y \mid C \oplus x; \mathcal{M})$$

where M' is fine-tuned on sample (x,y), C is context demonstrations

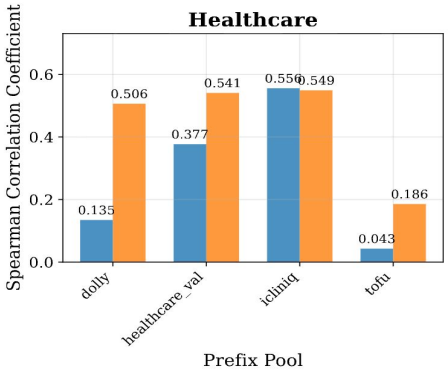
[1] Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers, (ACL 2023)
[2] On the Feasibility of In-Context Probing for Data Attribution (NAACL 2025)

From True Optimization to In-Context Approximation

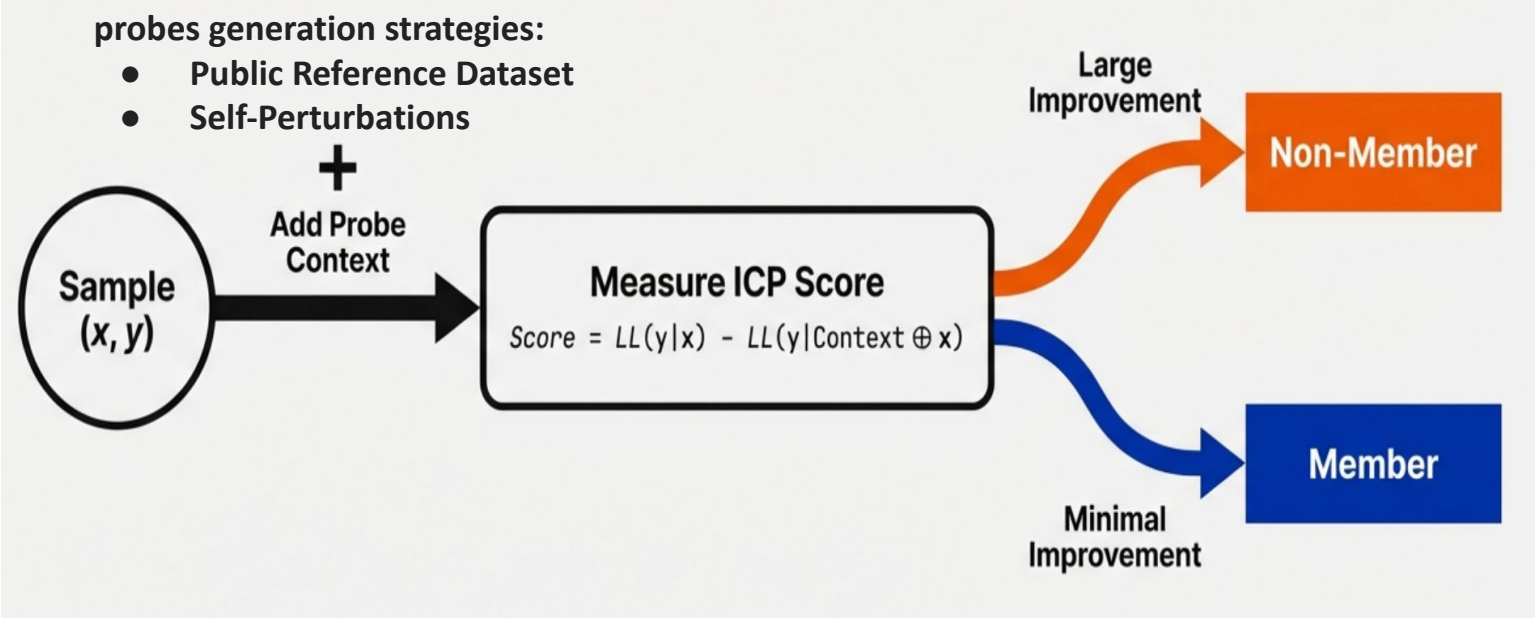
Empirical Study

We selected data from similar topics to the target dataset as probes and calculated the correlation between the actual fine-tuning loss reduction △LL and the simulation using in-context probes ICPscore.

- The more similar the topic and tasks, the higher the relevance.



ICP-MIA Framework



Evaluation Setup

ICP-MIA-SP Probes

C. Example of a masking-based context probe

x_{prompt} : Determine if the described symptoms relate to cystic fibrosis based on provided genetic information.
x_{input} : The patient exhibits regular bouts of persistent coughing, recurrent lung infections, and difficulty...
y : The described symptoms of regular bouts of persistent coughing, recurrent lung infections, and difficulty in...
C : The described symptoms of regular [MASK] of [MASK] [MASK] recurrent lung [MASK] and [MASK] in...

E. Prompt for perturbation generation

System:"You are a precise editor. Given the original text, generate a new text in which exactly 20 words are changed (added, removed, or replaced), but the overall meaning remains identical. Do not change more than 20 tokens. Output only the new text." **User:** "Original text:"

ICP-MIA-Ref Probes

D. Context Probe Example for ICP-MIA-REF

Instruction: If you are a doctor, please answer the medical questions based on the patient's description.
Question: I woke up this morning feeling the whole room is spinning when i was sitting down. I went to the bathroom walking unsteadily, as i tried to focus i feel nauseous. I try to vomit but it wont come out.. After taking panadol and sleep for few hours, i still feel the same....
Answer: Hi, Thank you for posting your query. The most likely cause for your symptoms is benign paroxysmal positional vertigo (BPPV), a type of peripheral vertigo. In this condition, the most common symptom is dizziness or giddiness, which is made worse with movements. ...

In-Context Probe:
Instruction: "If you are a doctor, please answer the medical questions based on the patient's description."
Question: "Hello doctor, After unsafe exposure, I got 49 days ELISA antibody test done, 71 days HIV proviral DNA PCR test, 87 days ELISA antibody test. All were negative. The antibody test I took is not a fourth generation test. Is it conclusive or should I take another test?...."
Answer: "Hi. Your tests are conclusive and you are not infected. No need for further tests. Ear ache and tongue papilla are not due to HIV and may be a simple bacterial infection...."

Evaluation Setup

Base Models: LLama3.2-3B, LLama3.2-3B-Instruct, Pythia-2.8B

Datasets: HealthCareMagic, MedInstruct, CNN-DM

Metrics: AUC, TPR@1%FPR

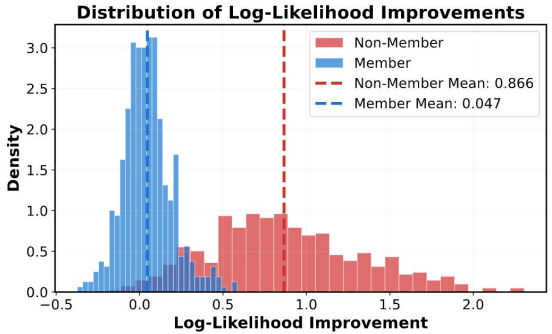
Evaluation Results

TABLE I: Comparison of MIA Methods across Different Models and Datasets

MIA Method	AUC								
	LLaMA-3.2-3B-Instruct			LLaMA-3.2-3B			Pythia-2.8B-deduped		
	Healthcare	MedInstruct	CNN-DM	Healthcare	MedInstruct	CNN-DM	Healthcare	MedInstruct	CNN-DM
Bag of Words	0.485	0.512	0.502	0.491	0.493	0.536	0.501	0.497	0.516
Loss Attack	0.770	0.907	0.929	0.708	0.904	0.885	0.701	0.849	0.851
Zlib	0.765	0.921	0.932	0.703	0.917	0.888	0.694	0.866	0.856
Min-K%	0.837	0.907	0.930	0.763	0.908	0.890	0.777	0.865	**0.859**
Min-K%++	0.798	0.810	0.861	0.710	0.787	0.794	0.727	0.758	0.760
Neighborhood	0.669	0.556	0.661	0.614	0.535	0.621	0.635	0.527	0.627
Recall	0.847	0.899	0.930	0.780	0.908	0.884	0.768	0.854	0.820
ICP-MIA-Ref	0.827	0.838	0.890	**0.842**	0.775	0.837	0.850	0.746	0.706
ICP-MIA-SP	**0.942**	**0.959**	**0.965**	0.763	**0.977**	**0.927**	**0.853**	**0.882**	0.845
Reference Attack (Base)*	0.796	0.885	0.925	0.736	0.878	0.871	0.717	0.871	0.856
Reference Attack (Ref)*	0.870	0.902	0.971	0.817	0.898	0.937	0.799	0.891	0.919
SPV-MIA*	0.781	0.946	0.974	0.725	0.932	0.959	0.713	0.869	0.938

MIA Method	TPR@1%FPR								
	LLaMA-3.2-3B-Instruct			LLaMA-3.2-3B			Pythia-2.8B-deduped		
	Healthcare	MedInstruct	CNN-DM	Healthcare	MedInstruct	CNN-DM	Healthcare	MedInstruct	CNN-DM
Bag of Words	0.008	0.014	0.004	0.014	0.010	0.002	0.003	0.002	0.008
Loss Attack	0.042	0.266	0.088	0.028	0.256	0.034	0.020	0.168	0.020
Zlib	0.036	0.096	0.058	0.034	0.076	0.042	0.006	0.028	0.034
Min-K%	0.046	0.288	0.104	0.024	0.412	0.090	0.022	0.176	0.046
Min-K%++	0.034	0.090	0.116	0.004	0.060	0.028	0.016	0.036	0.010
Neighborhood	0.032	0.008	0.012	0.014	0.010	0.008	0.018	0.010	0.006
Recall	0.024	0.133	0.195	0.014	0.044	0.096	0.020	0.162	0.108
ICP-MIA-Ref	0.084	0.044	0.020	**0.140**	0.018	0.062	0.110	0.074	0.022
ICP-MIA-SP	**0.172**	**0.326**	**0.518**	0.070	**0.538**	**0.418**	**0.122**	**0.270**	**0.144**
Reference Attack (Base)*	0.018	0.078	0.354	0.026	0.082	0.270	0.016	0.244	0.191
Reference Attack (Ref)*	0.012	0.166	0.388	0.010	0.142	0.412	0.010	0.414	0.390
SPV-MIA*	0.034	0.486	0.602	0.036	0.608	0.440	0.020	0.374	0.531

Note: **Bold** indicates the best overall performance, and underline indicates the second best performance among reference-free methods. For reference-based methods, Reference Attack (Base) uses the pretrained model itself as the reference, while Reference Attack (Ref) and SPV-MIA fine-tune a reference model on a held-out in-distribution split from the same dataset used to fine-tune the target model, giving them their strongest possible setting.



- Our ICP-MIA framework achieves better performance than all reference-free methods, and comparable to reference model based methods.
- ICP-MIA-SP has better performance in overall.
- Instruction-tuning tends to increase membership vulnerability.

Ablation Study on Probe Generation

TABLE II: ICP-MIA-SP Performance with Different Generator Models

Generator	Dataset	Llama3.2-3B		Llama3.2-3B-Instruct		Pythia-2.8B-Deduped	
		AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR
Qwen2.5-72B-Instruct	CNN-DM	**0.938**	**0.394**	0.968	0.533	0.750	0.020
	MedInstruct	**0.953**	0.176	**0.961**	**0.283**	**0.903**	0.077
	HealthcareMagic	0.792	**0.065**	**0.905**	0.098	**0.797**	**0.123**
Llama-3.3-70B-Instruct	CNN-DM	0.928	0.273	0.965	0.558	0.762	**0.025**
	MedInstruct	0.898	0.122	0.911	0.170	0.857	0.087
	HealthcareMagic	0.748	0.046	0.872	0.094	0.765	0.110
Mixtral-8x22B-Instruct	CNN-DM	0.933	0.382	0.966	**0.640**	0.748	0.018
	MedInstruct	0.940	0.108	0.947	0.108	0.901	**0.099**
	HealthcareMagic	0.758	0.060	0.898	0.057	0.787	0.080
GPT-4.1-mini	CNN-DM	0.920	0.124	**0.969**	0.340	**0.856**	0.010
	MedInstruct	0.940	**0.412**	0.946	0.260	0.876	0.054
	HealthcareMagic	**0.864**	0.042	0.850	**0.144**	0.735	0.016

- Stable performance with different generator models for probe generation

TABLE III: Public Dataset Impact on ICP-MIA-Ref

Model	Alpaca		iCliniq		TOFU		Validation	
	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
LLaMA-3.2-3B	0.813	0.072	**0.873**	0.188	0.743	0.020	0.864	**0.320**
LLaMA-3.2-3B-Instruct	0.819	0.168	0.857	0.112	0.821	0.146	**0.943**	**0.194**
Pythia-2.8B-Deduped	0.817	0.042	0.830	**0.122**	0.825	0.100	**0.875**	0.074

- Better attack performance with using more relevant public dataset as probes

MIA performance under DP-SGD

TABLE V: MIA AUC with different level DP-SGD

MIA Method	$\epsilon = 10$	$\epsilon = 50$	$\epsilon = 100$
Loss Attack	0.5080	0.5083	0.5121
Zlib	0.5012	0.5013	0.5064
Min-K%	0.5184	0.5189	0.5228
Min-K%++	0.5223	0.5225	0.5285
Neighborhood	0.5075	0.5080	0.5118
Recall	0.5150	0.5190	0.5232
Reference Attack(Ref)	0.5160	0.5200	0.5248
SPV-MIA	**0.5244**	0.5302	0.5332
ICP-MIA-SP	0.5235	**0.5310**	**0.5367**
ICP-MIA-REF	0.5112	0.5208	0.5244

- All attacks are mitigated by DP-SGD, our method still achieves better performance.

TABLE VI: MIA AUC in Label-Only Setting

MIA Method	Healthcare	MedInstruct	CNN-DM
PETAL	0.7354	0.7756	0.9018
ICP-MIA-Ref	0.7861	0.7632	0.7661
ICP-MIA-SP	**0.9143**	**0.7812**	**0.9351**

- Our method is adaptable to Text-Only Setting and better than baseline attack

Summary

1. We exploit a novel MIA signal —Optimization Gap—the disparity in remaining loss-reduction potential between member and non-member samples.

2. We Introduce In-Context Probing (ICP)-MIA, a training-free mechanism that simulates fine-tuning behavior at inference time, as attack vector.

3. Our experiments demonstrated that ICP-MIA outperforms existing reference-free attacks and achieves performance comparable to reference-model-based methods.