Was My Data Used for Training? Membership Inference in Open-Source LLMs via Neural Activations

Xue Tan[1], Hao Luan[1], Mingyu Luo[1], Zhuyang Yu[1], Jun Dai[2], Xiaoyan Sun[2], Ping Chen[1]

[1]Fudan University, [2]Worcester Polytechnic Institute

Background

- **Membership Inference Attacks (MIA)**

  Models generally exhibit higher confidence for training (member) samples than for unseen (non-member) samples, a phenomenon known as the *confidence gap*, which manifests in output probabilities, loss values, or attention weights.

- **Why MIA?**

  - Privacy verification: Detect misuse of private data in model training.
  - Compliance auditing: Audit data usage compliance (e.g., GDPR).
  - Copyright protection: Identify unauthorized use of copyrighted content.

Motivation

- **Why Open-Source LLMs?**
  - Widespread impact of open-source LLMs: Publicly available models hosted on the Hugging Face Hub exceed 900,000, with Meta's LLaMA family reaching over one billion downloads.
  - Scale and opacity of training data: The scale and opacity of training data challenge transparency, compliance, and data provenance.

- **Limitations of Existing Works**
  - **Black-box MIA:** Limited by output-only access, yielding low accuracy and high false positives.
  - **White-box MIA:** Modest gains (≈0.75 AUC) but high cost and poor scalability to long texts.

Contribution

- **Method**

  NART enables accurate membership inference by leveraging neural activations in open-source LLMs.

- **Benchmark**

  Construct three new benchmarks, WikiTection, NewsTection, and ArXivTection, using only post-cutoff data to enable rigorous membership inference evaluation.

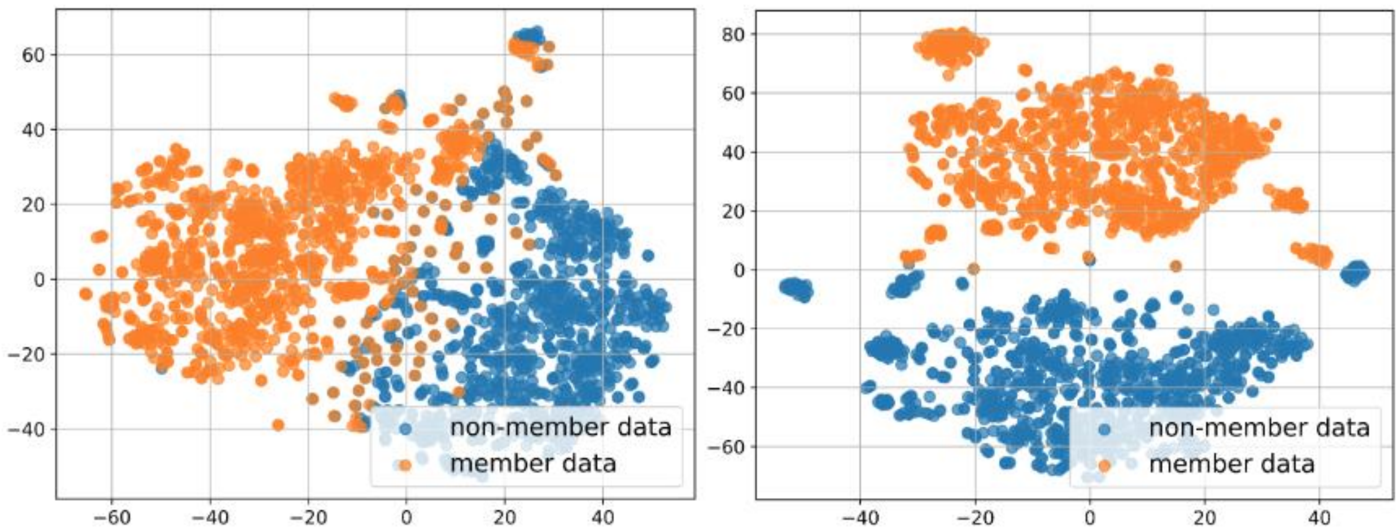
- **Evaluation**

  We demonstrate effectiveness across diverse architectures and challenging settings.

  (GPT-2, LLaMA2-7B , Mistral-7B, LLaMA3-8B, GPT-OSS-20B, Qwen3-8B, Pythia- 12B, and DCLM-1B; Pile , MIMIR, and DCLM)

Motivation Example

- **Activation distinguishable**



(a) LLaMA3-8B (b) Mistral-7B

- **Challenges**
  - Challenge1:Data overlap and leakage
    - Overlap between training data and public datasets
  - Challenge2:Activation feature challenges
    - High-dimensional
    - Subtle activation differences
    - Variable-length inputs
  - Challenge3: Limited supervision
    - Practical settings require effective inference under limited supervision

Design

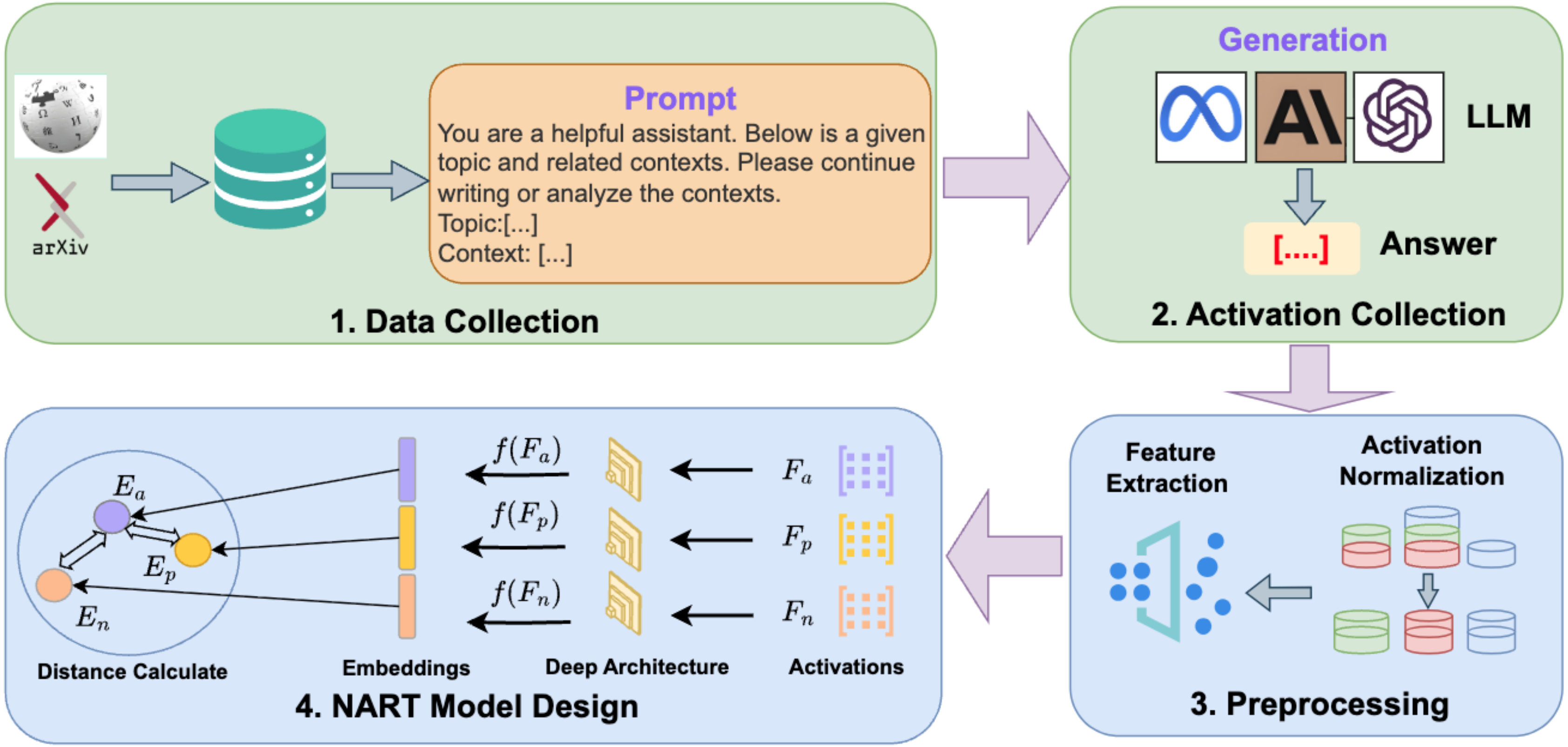


Fig. 2: The workflow of NART.

Method

- **Data Collection**

  - WikiTection, NewsTection, and ArXivTection.
  - Dataset is split into member and non-member data, with the model fine-tuned on the member set.

- **Activation Collection**

  - Divide input text D into N_s subsequences S_j of length C.
  - Get the activation of the last token.
  - Normalize the activation.

$$Nor_j = (Act_j - \mu) / \sigma.$$

Method

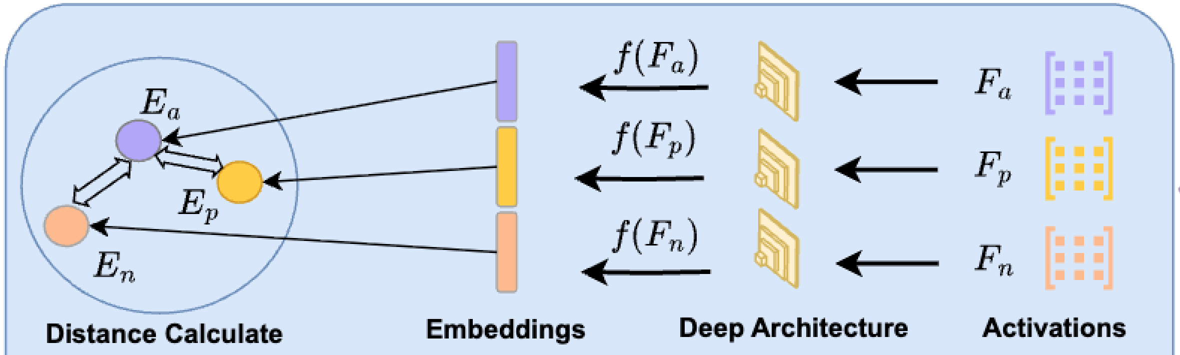
- **Activation Preprocessing**

  - ***NonFE:*** Directly use the normalized activation Nor_j of the last token in subsequence S_j as the feature representation of that subsequence
  - ***StatFE:*** Compute statistical features from each layer of the normalized activations, including the minimum, maximum, mean, and standard deviation
  - ***HistFE:*** characterizes the activation distribution of subsequence S_j in D by constructing a histogram.

Method

● NART Model Design

- ➢ Learns a discriminative activation distance space using a Siamese-based triplet network.
- ➢ ResNet18 is adopted to model cross-layer and intra-layer activation relationships.
- ➢ Contrastive learning is used to amplify member–non-member differences under few-shot settings



4. NART Model Design

Method

- **Model Training**

  - Contrastive Learning: Model training is based on a **triplet** data structure, where each triplet consists of three components ($\{F_a, F_p, F_n\}$).
    - The **anchor** F_a serves as the reference sample.
    - The **positive** sample F_p belongs to the same class as the anchor (e.g., both are member data).
    - The **negative** sample F_n belongs to a different class.
  - Loss function:

$$L = \max(\mathrm{Dist}\left(E_a, E_p\right) - \mathrm{Dist}\left(E_a, E_n\right) + \alpha, 0)$$

Method

Inference via Support-Set Voting

- Subsequence embedding: A long document D_t is segmented into subsequences S_j, each mapped to an embedding E_{S_j}
- Nearest-neighbor labeling: Each subsequence is assigned the label of its nearest support sample in the embedding space.
- Majority voting: The document label is determined by majority voting over subsequences.

$$L(D_t) = \mathbb{1}\left[\sum_{j=1}^{N_s} L_{S_j} \geq \frac{N_s}{2}\right]$$

Evaluation

- **Model**
  - GPT-2, LLaMA2-7B , Mistral-7B, LLaMA3-8B, GPT-OSS-20B, Qwen3-8B, Pythia-12B, and DCLM-1B
- **Dataset**
  - WikiTection, NewsTection, ArXivTection, Pile, MIMIR, and DCLM
- **Baseline**
  - Black-box: *Loss attack, Zlib, Lowercase, Min-K% Prob, Neighborhood*
  - White-box: *PARSING, Probe*

Dataset	**Target Model**			
	learning rate	**batch size**	**epoch**	**max-seq-len**
WikiTection	2e-5	16	3	512
NewsTection	2e-5	16	4	512
ArxivTection	2e-5	16	3	2048

Evaluation

- **Overall Results**

Dataset	FeaEXTRACT	Metrics	*LLMs*						
			GPT2-xl	LLaMA2-7B	LLaMA3-8B	Mistral-7B	LLaMA2-13B	GPT-OSS-20B	Qwen3-8B
WikiTection	***NonFE***	TPR@5%FPR	0.991	0.941	0.991	0.997	0.989	0.990	0.996
		AUC	0.992	0.981	0.998	0.996	0.997	0.994	0.999
	StatFE	TPR@5%FPR	0.986	0.991	0.996	0.992	0.987	0.988	0.990
		AUC	0.997	0.994	0.991	0.998	0.997	0.993	0.992
	HistFE	TPR@5%FPR	0.995	0.982	0.991	0.995	0.991	0.990	0.993
		AUC	0.995	0.996	0.995	0.997	0.996	0.994	0.998
NewsTection	***NonFE***	TPR@5%FPR	0.918	0.900	0.941	0.959	0.982	0.972	0.984
		AUC	0.977	0.976	0.988	0.985	0.986	0.980	0.988
	StatFE	TPR@5%FPR	0.982	0.972	0.918	0.977	0.964	0.980	0.978
		AUC	0.994	0.988	0.971	0.996	0.989	0.989	0.984
	HistFE	TPR@5%FPR	0.982	0.941	0.939	0.973	0.964	0.986	0.990
		AUC	0.988	0.985	0.965	0.991	0.984	0.990	0.994
ArXivTection	***NonFE***	TPR@5%FPR	0.973	0.941	0.988	0.982	0.977	0.984	0.982
		AUC	0.986	0.967	0.994	0.989	0.993	0.990	0.986
	StatFE	TPR@5%FPR	0.982	0.941	0.982	0.973	0.991	0.990	0.988
		AUC	0.998	0.984	0.993	0.994	0.996	0.994	0.990
	HistFE	TPR@5%FPR	0.986	0.961	0.982	0.964	0.996	0.990	0.992
		AUC	0.991	0.985	0.997	0.992	0.999	0.995	0.997

Evaluation

- **Robustness**

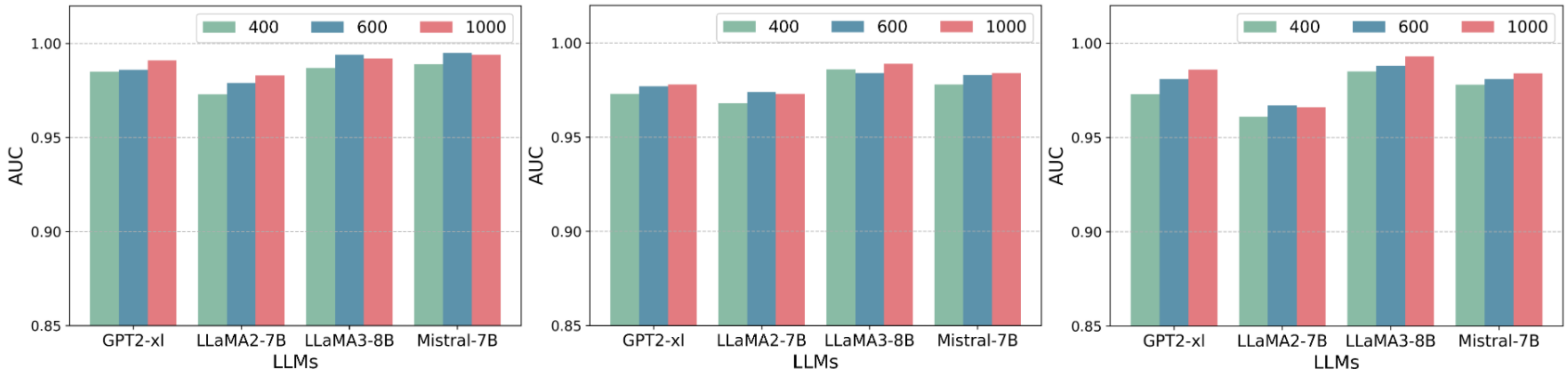
Paraphrased Dataset	Metrics	*LLMs*			
		GPT2-xl	LLaMA2-7B	LLaMA3-8B	Mistral-7B
WikiTection_para	TPR@5%FPR	0.978	0.942	0.969	0.987
	AUC	0.989	0.973	0.995	0.997
NewsTection_para	TPR@5%FPR	0.841	0.827	0.841	0.959
	AUC	0.969	0.953	0.965	0.988
ArXivTection_para	TPR@5%FPR	0.971	0.934	0.982	0.971
	AUC	0.983	0.968	0.997	0.989

Dataset	Mislabeled Ratio	GPT2-xl		LLaMA3-8B		Mistral-7B	
		AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR
WikiTection	5%	0.988	0.941	0.986	0.961	0.939	0.686
	10%	0.956	0.757	0.923	0.757	0.872	0.398
	20%	0.861	0.486	0.863	0.438	0.768	0.181
NewsTection	5%	0.982	0.961	0.945	0.853	0.968	0.951
	10%	0.971	0.893	0.894	0.427	0.940	0.748
	20%	0.946	0.733	0.808	0.405	0.818	0.259
ArXivTection	5%	0.991	0.941	0.989	0.971	0.970	0.951
	10%	0.962	0.874	0.959	0.709	0.938	0.689
	20%	0.814	0.143	0.829	0.171	0.848	0.295

Evaluation

- **Generation**

Dataset	Dataset Size	GPT2-xl AUC	GPT2-xl TPR@5%FPR	LLaMA3-8B AUC	LLaMA3-8B TPR@5%FPR	Mistral-7B AUC	Mistral-7B TPR@5%FPR
WikiTection	50	0.994	0.989	0.991	0.976	0.995	0.993
	100	0.994	0.990	0.998	0.989	0.996	0.990
	200	0.995	0.991	0.999	0.990	0.998	0.996
ArXivTection	50	0.988	0.973	0.998	0.987	0.990	0.986
	100	0.991	0.978	0.996	0.989	0.988	0.989
	200	0.989	0.980	0.993	0.991	0.991	0.983



(a) WikiTection (b) NewsTection (c) ArXivTection

Evaluation

- **Generation**

Imbalanced Training Data

TABLE IV: Evaluating NART under imbalanced datasets.

Testing Dataset	Metrics	*LLMs*			
		GPT2-xl	LLaMA2-7B	LLaMA3-8B	Mistral-7B
WikiTection	TPR@5%FPR	0.986	0.934	0.983	0.975
	TPR@10%FPR	0.988	0.965	0.991	0.980
	AUC	0.989	0.976	0.989	0.986
NewsTection	TPR@5%FPR	0.911	0.902	0.938	0.955
	TPR@10%FPR	0.932	0.965	0.973	0.969
	AUC	0.973	0.977	0.985	0.981
ArXivTection	TPR@5%FPR	0.971	0.940	0.988	0.975
	TPR@10%FPR	0.979	0.956	0.990	0.979
	AUC	0.978	0.966	0.989	0.981

Training and Testing Data from Different Sources

TABLE XIII: Generalization performance of NART. Abbreviations: Wiki (WikiTection), News (NewsTection), ArXiv (ArXivTection).

Training Dataset	Test Dataset	Metrics	LLMs		
			GPT2-xl	LLaMA3-8B	Mistral-7B
Wiki & News	ArXiv	TPR@5%FPR	0.567	0.405	0.894
		AUC	0.918	0.751	0.970
News & ArXiv	Wiki	TPR@5%FPR	0.932	0.959	0.951
		AUC	0.972	0.989	0.978
Wiki & ArXiv	News	TPR@5%FPR	0.206	0.526	0.209
		AUC	0.712	0.889	0.762

Evaluation

- **Performance on Pretrained Language Models**

TABLE XV: The performance of NART on Pythia-12B.

Dataset	FeaEXTRACT	Pythia-12B				
		AUC	**TPR@10%FPR**	**TPR@5%FPR**	**TPR@3%FPR**	**TPR@1%FPR**
Pile	***NonFE***	0.923	0.911	0.683	0.475	0.188
	StatFE	0.931	0.916	0.691	0.415	0.201
	HistFE	0.918	0.899	0.654	0.489	0.199
MIMIR	***NonFE***	0.955	0.931	0.802	0.446	0.305
	StatFE	0.932	0.921	0.525	0.465	0.297
	HistFE	0.967	0.941	0.832	0.723	0.356

TABLE XVI: The performance of NART on DCLM-1B.

Dataset	FeaEXTRACT	DCLM-1B				
		AUC	**TPR@10%FPR**	**TPR@5%FPR**	**TPR@3%FPR**	**TPR@1%FPR**
DCLM-Baseline	***NonFE***	0.974	0.980	0.970	0.584	0.386
	StatFE	0.935	0.901	0.733	0.446	0.149
	HistFE	0.988	0.990	0.980	0.842	0.436

Evaluation

- **Efficiency**

TABLE VI: Training and inference time comparisons across datasets and models.

Model	Training Time per Epoch			Inference Time per Sample		
	WikiTection	NewsTection	ArXivTection	WikiTection	NewsTection	ArXivTection
PARSING	11.15s	10.76s	10.89s	0.0019s	0.0015s	0.0018s
NART_NonFE	18.34s	18.52s	18.62s	0.0060s	0.0059s	0.0061s
NART_StatFE	0.58s	0.56s	0.58s	0.0054s	0.0034s	0.0035s
NART_HistFE	1.99s	1.21s	1.66s	0.0098s	0.0098s	0.0056s

- **Blind Attacks on Datasets**

Dataset	Date Detection		Bag-of-words classification		Greedy rare word selection	
	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR
WikiTection	0.512	0.056	0.514	0.066	0.495	0.075
NewsTection	0.497	0.051	0.489	0.044	0.526	0.086
ArXivTection	0.523	0.086	0.492	0.045	0.502	0.060

Conclusion

Thanks!