Fine-tuning can Help Detect Pretraining Data from Large Language Models

Hengxiang Zhang Songxin Zhang Bingyi Jing Hongxin Wei

ICLR 2025

Pretraining Data Detection

Pretraining data detection can be formulated as a binary classification: determining whether a given text x is a member or non-member of the pretraining dataset. A level-set estimation can perform pretraining data detection:

$$h(\boldsymbol{x}; f_{\boldsymbol{\theta}}) = \begin{cases} \text{member} & \text{if } \mathcal{S}(\boldsymbol{x}; f_{\boldsymbol{\theta}}) < \epsilon, \\ \text{non-member} & \text{if } \mathcal{S}(\boldsymbol{x}; f_{\boldsymbol{\theta}}) \geq \epsilon, \end{cases}$$

where $S(x; f_\theta)$denotes a scoring function, and ϵ is the threshold determined by a validation dataset.

Scoring Functions

- **Perplexity** is proposed to distinguish members and non-members, based on the observation that members tend to have lower perplexity than non-members. The perplexity of x is calculated as:

$$\text{Perplexity}(\boldsymbol{x}; f_{\boldsymbol{\theta}}) = \exp\{-\frac{1}{n}\sum_{i=1}^{n}\log p_{\boldsymbol{\theta}}(x_i \mid x_1, \ldots, x_{i-1})\}$$

where $x = \{x_1, x_2, \ldots, x_n\}$ is a sequence of tokens and $p_\theta, (x_i \mid x_1, \ldots, x_{i-1})$ is the conditional probability of x_i given the preceding tokens.

Scoring Functions

- **Min-k%** computes the average probabilities of k% outlier tokens with the smallest predicted probability. The intuition is that a nonmember example is more likely to include a few outlier words with low likelihoods than members. Min-k% is computed by:

$$\text{Min-k\%}(\boldsymbol{x}; f_{\boldsymbol{\theta}}) = \frac{1}{E} \sum_{x_i \in \text{Min-k\%}(\boldsymbol{x})} \log p_{\boldsymbol{\theta}}(x_i \mid x_1, \ldots, x_{i-1})$$

where E is the size of the Min-k%(x) set.

Unsatisfactory Performance

- Non-member data can obtain low perplexities by including frequent or repetitive texts, while members may contain rare tokens that result in high perplexities.

- The significant overlap in scores distribution between members and non-members makes it hard to distinguish between them.

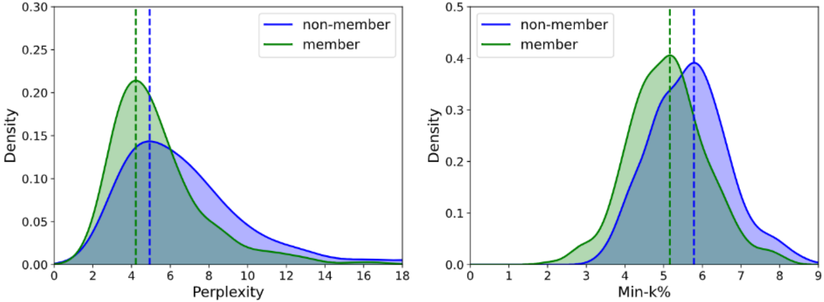


Figure: The scores distribution of perplexity and Min-k% from the pre-trained model.

Motivation: Fine-tuning with Nonmembers

- Unseen data in the pretraining tend to obtain a lower perplexity from the fine-tuned model than the pre-trained model.

- The shift in perplexity distribution for members is negligible after fine-tuning.

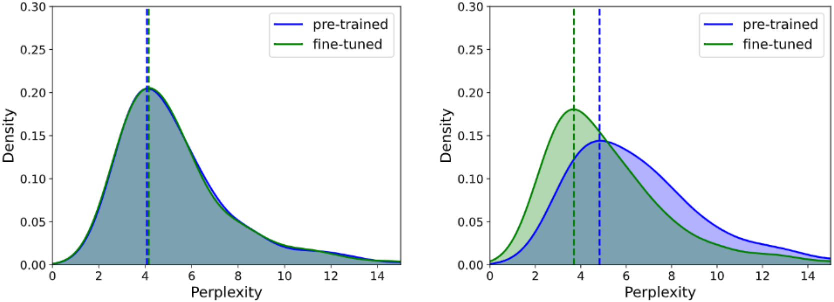


Figure: The perplexity distribution from the pre-trained model and the fine-tuned model.

Method: Fine-tuned Score Deviation (FSD)

- **Fine-tuned Score Deviation** is proposed to exploit the score deviation for detecting pretraining data. Given a sample x, we calculate the score difference between the pre-trained LLM f_θ and the fine-tuned LLM $f_{\theta'}$. The new score is formulated as:

$$\mathrm{FSD}(\boldsymbol{x}; f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'}) = \mathcal{S}(\boldsymbol{x}; f_{\boldsymbol{\theta}}) - \mathcal{S}(\boldsymbol{x}; f_{\boldsymbol{\theta}'})$$

where θ' denotes the parameters of LLM after fine-tuning, and $S(\cdot)$ denotes an existing scoring function, such as Perplexity and Min-k%.

Procedure of Our Method

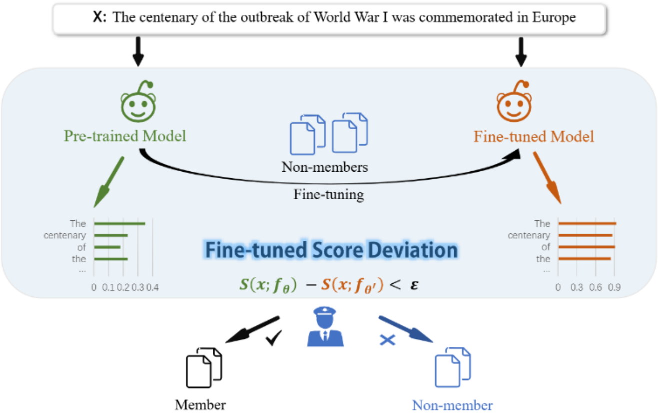


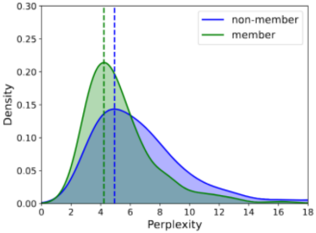
Figure: Overview of Fine-tuned Score Deviation.

I. Collect a small amount of unseen data for the LLM within the same domain.

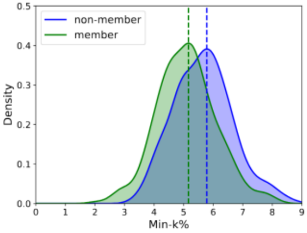
II. Perform fine-tuning on LLMs with the constructed fine-tuning dataset.

III. Calculate the score difference between the pre-trained and fine-tuned LLM.

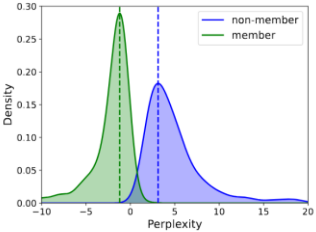
Clear Distinction with FSD



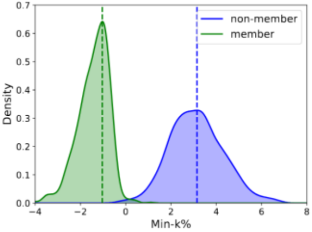
(a) Perplexity



(b) Min-k%



(c) FSD with Perplexity



(d) FSD with Min-k%

Results on WikiMIA and ArXivTection

Table: AUC score for pretraining data detection with baselines and our method from various models.

Dataset	Method	GPT-J-6B		OPT-6.7B		Pythia-6.9B		LLaMA-7B		NeoX-20B	
		Base	*+Ours*	*Base*	*+Ours*	*Base*	*+Ours*	*Base*	*+Ours*	*Base*	*+Ours*
WikiMIA	Perplexity	0.64	**0.95**	0.60	**0.90**	0.64	**0.90**	0.64	**0.92**	0.69	**0.93**
	Lowercase	0.59	**0.77**	0.59	**0.71**	0.58	**0.74**	0.58	**0.69**	0.66	**0.76**
	Zlib	0.61	**0.94**	0.59	**0.89**	0.61	**0.88**	0.62	**0.90**	0.64	**0.93**
	MIN-K%	0.68	**0.92**	0.62	**0.91**	0.67	**0.86**	0.65	**0.85**	0.73	**0.90**
ArXivTection	Perplexity	0.79	**0.96**	0.68	**0.89**	0.77	**0.95**	0.68	**0.92**	0.79	**0.95**
	Lowercase	0.59	**0.81**	0.58	**0.70**	0.60	**0.77**	0.50	**0.69**	0.62	**0.75**
	Zlib	0.64	**0.96**	0.55	**0.89**	0.63	**0.95**	0.57	**0.91**	0.65	**0.95**
	MIN-K%	0.85	**0.92**	0.74	**0.84**	0.84	**0.91**	0.76	**0.86**	0.85	**0.91**

- FSD significantly improves the performance of all baselines across diverse models.

Results on the Pile Dataset

Table: The average AUC score of baselines and our method from the Pythia-6.9B over 20 subsets of the Pile dataset.

Method	**Perplexity**		**Lowercase**		**Zlib**		**MIN-K%**	
	Base	***+Ours***	*Base*	***+Ours***	*Base*	***+Ours***	*Base*	***+Ours***
Pile	0.503	**0.625**	0.519	**0.566**	0.507	**0.624**	0.515	**0.600**

- Our FSD improves the performance of baselines on the Pile dataset under the Pythia-6.9B model.

Results on Varying the Fine-tuning Data Size

- Our FSD can improve the performance of baselines with a few non-members, demonstrating its practicality.

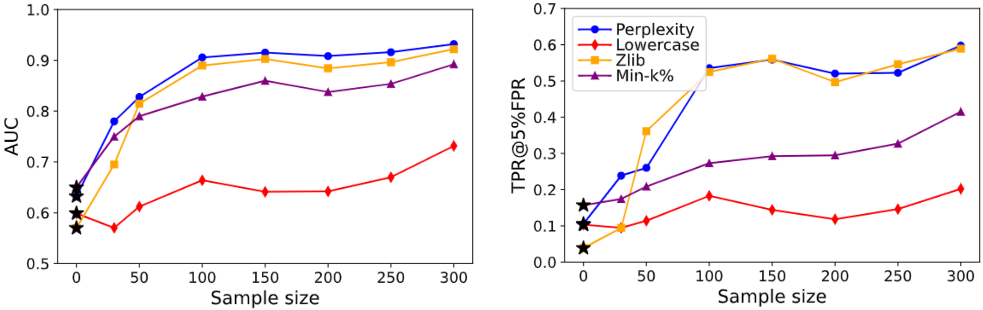


Figure: AUC and TPR@5%FPR of FSD, using auxiliary datasets with varying sizes.

Results on Copyrighted Book Detection

Table: Accuracy and AUC score for copyrighted book detection.

Metric	Accuracy				AUC			
Method	BookTection		BookMIA		BookTection		BookMIA	
	Base	***+Ours***	*Base*	***+Ours***	*Base*	***+Ours***	*Base*	***+Ours***
Perplexity	66.9	**85.4**	59.0	**96.5**	0.710	**0.910**	0.564	**0.995**
Lowercase	64.5	**73.0**	67.0	**69.2**	0.664	**0.770**	0.708	**0.779**
Zlib	65.3	**86.4**	57.4	**98.6**	0.568	**0.920**	0.474	**0.999**
MIN-K%	68.1	**82.1**	59.5	**93.9**	0.716	**0.880**	0.587	**0.979**

- Our FSD significantly improves the accuracy of baseline methods for copyrighted book detection.

Ablation Study

Table: AUC and TPR@5%FPR of FSD with different fine-tuning methods.

Metric	AUC				TPR@5%FPR			
Method	Base	AdaLoRA	IA3	LoRA	Base	AdaLoRA	IA3	LoRA
Perplexity	0.64	0.82	0.91	**0.92**	0.09	0.39	**0.52**	0.41
Lowercase	0.58	0.62	**0.72**	0.69	0.10	0.13	0.17	**0.18**
Zlib	0.62	0.76	0.84	**0.90**	0.09	0.24	0.32	**0.47**
MIN-K%	0.65	0.80	**0.90**	0.85	0.15	0.22	**0.39**	0.25

- Our FSD can be implemented with different fine-tuning methods and does not require a specific finetuning technique.

Conclusion

- **Challenge:** Unseen data can obtain high likelihood by including frequent or repetitive texts, while seen data may contain rare tokens that result in low likelihood, which casts a challenge for detecting pretraining data.

- **Motivation:** Compared to non-member data, member data experience a greater perplexity shift after fine-tuning with a few non-members.

- **Method:** Fine-tuned Score Deviation (FSD) is proposed to measure the deviation distance of current scores after fine-tuning on a small amount of unseen data within the same domain.

Paper: https://openreview.net/pdf?id=X8dzvdkQwO

Code: https://github.com/ml-stat-Sustech/Fine-tuned-Score-Deviation