Rethinking Pretraining Data Detection for LLMs: From Local to Global

Chenye Ke[1], Yan Zhuang[2], Zirui Liu[1], Qi Liu[1]*

[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
[2]Nanjing University of Aeronautics and Astronautics

The 43rd International Conference on Machine Learning

May 2026

Overview

- Statement of the Problem
- Motivation
- Core Insight
- The Proposed AECA
- Experiments
- Analysis
- Conclusion

Statement of the Problem

☐ What is pretraining data detection?

Infer whether a target text was included in an LLM's pretraining corpus.

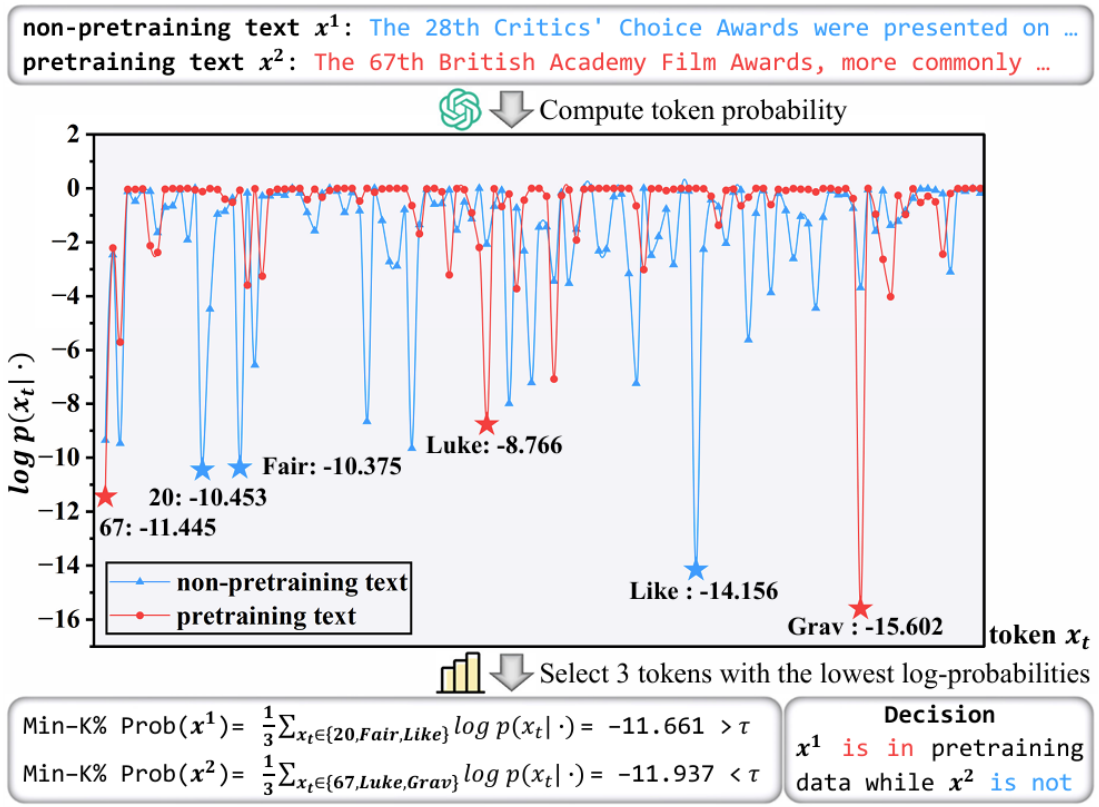
☐ Why does this task matter?

It helps audit privacy leakage and benchmark contamination.

☐ What access does the detector have?

Gray-box setting: use output statistics such as loss, logits, and token probabilities - not weights or the corpus.

Motivation



Existing methods

- select a few individual tokens which the model finds surprising
- aggregate static local statistics
- ignore probability dynamics across the whole sequence

$$s(x;\mathcal{M}) = \frac{1}{|E(x)|}\sum_{x_t\in E(x)} F(\log p(x_t|x_{<t}))$$

Key limitation:

Local minima do not necessarily represent the model's holistic prediction state.

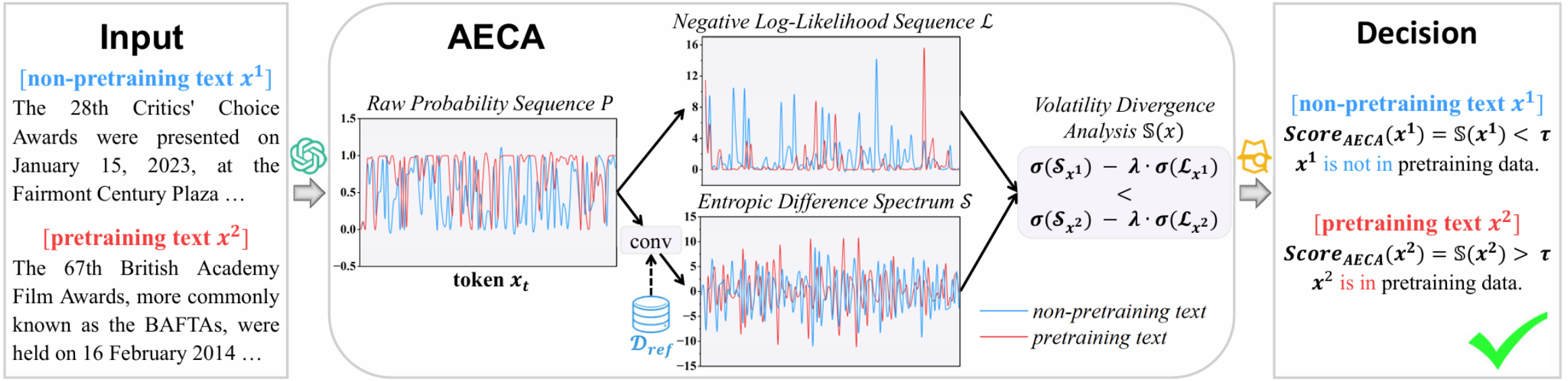
Core Insight

Our Core Idea

Treat the probability sequence as a dynamic signal.

- memorized sequences show more deterministic retrieval patterns
- generalized generation involves uncertainty and fluctuation
- the detection signal lies in volatility divergence

AECA (I)



1. **Probability Acquisition:** get raw token probabilities from the LLM.
2. **Adaptive Entropic Convolution:** calibrate probabilities and apply convolution to amplify fluctuations.
3. **Volatility Divergence Analysis:** calculate volatility difference between entropy and NLL sequences.
4. **Decision:** classify text based on the volatility divergence score.

AECA (II)

Adaptive Entropic Convolution: Uncover and highlight memorization signals

$$P_{\text{ref}}(x_t) = \frac{C(x_t; \mathcal{D}_{\text{ref}}) + \alpha}{N + \alpha|\mathcal{V}|} \Rightarrow \mathcal{I}_{\text{self}}(x_t) = -\log P_{\text{ref}}(x_t) \Rightarrow \Phi(t) \triangleq p_\theta(x_t \mid x_{<t}) \cdot \mathcal{I}_{\text{self}}(x_t)$$

$$\Downarrow$$

$$\mathcal{S}[t] = (\Phi * \mathbf{k})[t] \triangleq \Phi(t) - \Phi(t+1)$$

Why calibration? [Definition 4.1]

- ✓ Common words are easy and noisy.
- ✓ Rare-token high confidence better exposes memorization.
- ✓ Couples model confidence with token scarcity.

Why convolution? [Proposition 4.3]

- ✓ Treats Φ as a sequence signal.
- ✓ Suppresses smooth low-frequency trends.
- ✓ Amplifies abrupt token-to-token changes.

Zhang W, et al. Pretraining data detection for large language models: A divergence-based calibration method[C]//EMNLP 2024.

AECA (III)

Volatility Divergence Analysis: Compute the divergence score for final detection

$$\mathbf{S}_{\mathrm{AECA}}(x) = \sigma(\mathcal{S}_x) - \lambda \cdot \sigma(\mathcal{L}_x)$$

[Lemma 4.4] There exists a coefficient λ such that AECA assigns higher scores to memorized texts:

$$\lambda > \frac{(2\gamma+\epsilon)\log\frac{N+\alpha|\mathcal{V}|}{\alpha}+\log\frac{N+\alpha}{\alpha}}{\mu-(\gamma+\frac{\epsilon}{2})}$$

$$[\sigma(\mathcal{S}_x) - \lambda\sigma(\mathcal{L}_x)]_{\mathbb{M}} > [\sigma(\mathcal{S}_x) - \lambda\sigma(\mathcal{L}_x)]_{\mathbb{G}}$$

Key intuition:

Memorized samples: NLL volatility tends to collapse, while entropic difference volatility is preserved. Detailed proof is provided in Appendix B of the paper.

Experiments - WikiMIA Results

Len.	Method	Mamba-1.4B		GPT-Neo-2.7B		OPT-6.7B		Pythia-6.9B		Pythia-12B		GPT-NeoX-20B		Avg.
		Ori.	*Para.*	*Ori.*	*Para.*	*Ori.*	*Para.*	*Ori.*	*Para.*	*Ori.*	*Para.*	*Ori.*	*Para.*	
128	PPL	0.636	0.631	0.636	0.632	0.622	0.613	0.659	0.657	0.660	0.659	0.699	0.689	0.649
	Ref	0.599	0.609	0.585	0.561	0.632	0.637	0.664	0.662	0.657	0.658	0.680	0.631	0.631
	Lowercase	0.584	0.591	0.601	0.604	0.580	0.586	0.608	0.619	0.619	0.624	0.658	0.670	0.612
	Zlib	0.658	0.612	0.662	0.611	0.640	0.598	0.681	0.635	0.682	0.638	0.717	0.666	0.650
	Min-K%	0.664	0.676	0.664	0.673	0.663	0.665	0.700	0.701	0.706	0.713	0.742	0.737	0.692
	Min-K%++	0.672	0.676	0.652	0.649	0.674	0.680	0.696	0.697	**0.722**	0.724	0.691	0.702	0.686
	DC-PDD	0.666	0.666	0.669	0.668	0.670	0.671	0.688	0.701	0.688	0.704	0.746	**0.752**	0.699
	PAC	0.673	0.683	0.647	0.682	0.618	0.641	0.709	0.714	0.710	0.730	0.735	0.742	0.690
	AECA	**0.682**	**0.692**	**0.698**	**0.696**	**0.686**	**0.688**	**0.718**	**0.726**	0.716	**0.734**	**0.751**	**0.752**	**0.712**
256	PPL	0.666	0.637	0.674	0.628	0.639	0.596	0.692	0.658	0.688	0.653	0.713	0.697	0.662
	Ref	0.584	0.614	0.623	0.670	0.691	0.634	0.649	0.694	0.646	0.691	0.686	0.702	0.657
	Lowercase	0.588	0.622	0.629	0.634	0.612	0.611	0.592	0.633	0.633	0.653	0.635	0.655	0.625
	Zlib	0.679	0.526	0.686	0.518	0.662	0.505	0.703	0.553	0.701	0.548	0.729	0.575	0.615
	Min-K%	0.701	0.684	0.706	0.692	0.669	0.633	0.710	0.700	0.721	0.683	0.743	0.719	0.697
	Min-K%++	0.637	0.603	0.677	0.623	0.626	0.595	0.608	0.617	0.643	0.600	0.598	0.601	0.619
	DC-PDD	0.626	0.618	0.670	0.658	0.625	0.598	0.664	0.682	0.687	0.677	0.737	0.680	0.660
	PAC	0.689	0.705	0.677	0.679	0.658	0.646	0.711	**0.728**	0.716	0.691	0.724	0.709	0.694
	AECA	**0.707**	**0.710**	**0.715**	**0.708**	**0.696**	**0.662**	**0.713**	0.727	**0.729**	**0.698**	**0.758**	**0.748**	**0.714**

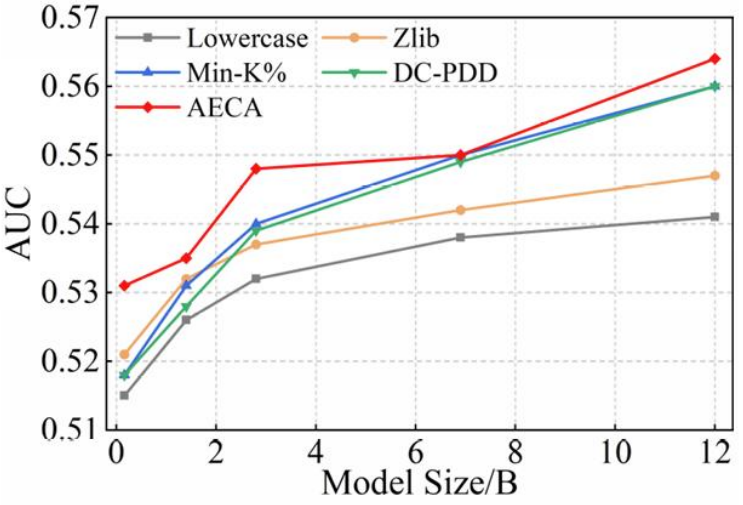
- ✓ **Long-Text Advantage:** **+1.3 / +1.7** AUC points over the best baseline
- ✓ **Robustness:** Strong on paraphrased texts and non-Transformer architectures

Experiments - MIMIR Results

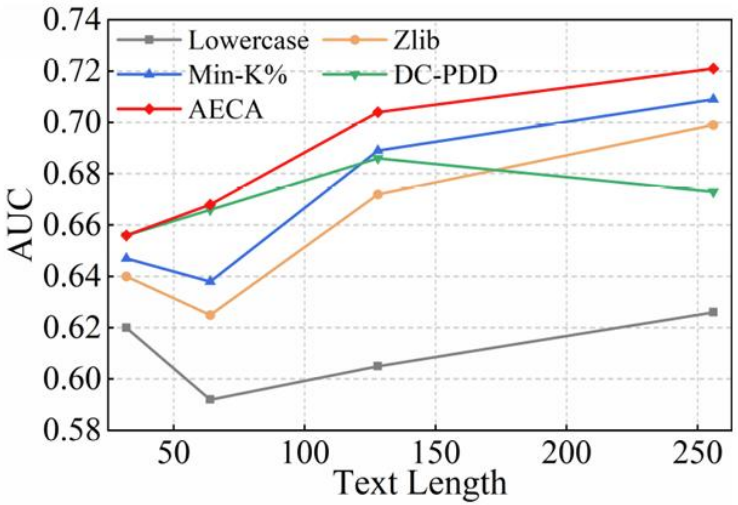
Method	ArXiv					HackerNews					PubMed Central					Avg.
	160M	1.4B	2.8B	6.9B	12B	160M	1.4B	2.8B	6.9B	12B	160M	1.4B	2.8B	6.9B	12B	
PPL	0.543	**0.558**	0.564	**0.573**	0.579	0.502	0.517	0.525	0.532	0.539	0.507	0.520	0.526	0.533	0.537	0.537*
Ref	0.494	0.532	0.539	0.553	0.563	0.488	0.518	0.535	**0.546**	**0.558**	0.498	0.524	0.530	0.537	0.541	0.530*
Lowercase	0.529	0.543	0.552	0.558	0.566	0.494	0.508	0.517	0.525	0.523	0.523	0.526	0.527	0.531	0.534	0.530*
Zlib	0.538	0.552	0.557	0.565	0.571	0.511	0.519	0.524	0.527	0.532	0.513	0.525	0.530	0.535	0.539	0.536*
Min-K%	0.530	0.549	0.559	0.572	0.582	0.512	0.519	0.530	0.540	0.553	0.511	0.524	0.531	0.537	0.545	0.540†
Min-K%++	0.521	0.543	0.559	0.568	0.576	0.511	0.515	0.527	0.542	0.565	0.513	0.522	0.529	0.542	**0.551**	0.539*
DC-PDD	0.528	0.549	0.557	0.563	0.575	0.513	0.513	0.530	0.542	0.552	0.513	0.521	0.529	0.543	0.553	0.539*
PAC	0.527	0.549	0.562	0.572	0.582	0.485	0.492	0.513	0.519	0.532	0.510	0.526	0.535	**0.546**	**0.551**	0.533†
AECA	**0.548**	0.553	**0.569**	0.571	**0.587**	**0.518**	**0.523**	**0.531**	0.541	**0.558**	**0.526**	**0.528**	**0.544**	0.539	0.547	**0.546**

- ✓ **Same-Source Robustness:** +0.6 AUC points over the best baseline
- ✓ **Cross-Domain Generalization:** Consistent gains across domains and model scales

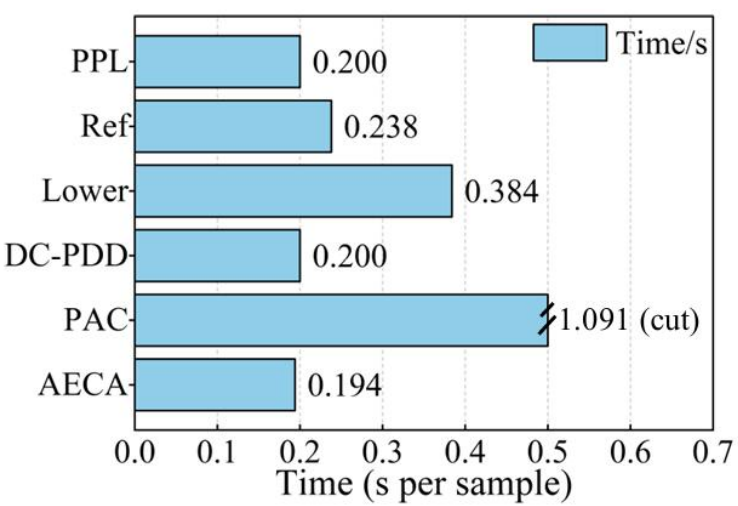
Analysis



(a) AUC score vs. model size



(b) AUC score vs. text length



(c) computational efficiency

- ✓ **Scalability:** Consistent gains across model sizes
- ✓ **Length Robustness:** Larger advantage on longer texts
- ✓ **Efficiency:** 0.194 s/sample, much faster than PAC

AECA is both performant and efficient, making it practical for large-scale detection.

Conclusion

- We **reframe** pretraining data detection from local token statistics to global sequence dynamics.
- AECA integrates self-information calibration, entropic convolution, and volatility divergence with **theoretical guarantees**.
- It is accurate, robust, and efficient - especially for **long-text** detection.

Q&A Session

Thank you for your attention!

[AECA Code]

[AECA Poster]

Contact: kechenye03@gmail.com