# Enhancing Pre-Training Data Detection through Distribution Shape Analysis: A Multi-Scale Weighted Residual Approach to Min-K%++

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Pre-training data detection in large language models has emerged as a critical challenge for model transparency and compliance, with membership inference attacks serving as the primary mechanism for identifying whether specific text sequences were part of a model's training data. While Min-K%++ represents the current state-of-the-art approach, it suffers from a fundamental limitation: uniform aggregation of token-level scores ignores valuable distributional patterns that could enhance detection accuracy. We propose a novel enhancement through residual score decomposition with multi-scale importance weighting, which analyzes distribution shape features such as skewness, kurtosis, and entropy to reveal training versus non-training patterns. Our method decomposes Min-K%++ scores into trend and residual components using exponential moving averages (Lucas & Saccucci, 1990), applies position-based weighting that emphasizes earlier tokens in sequences, and performs multi-scale deviation analysis to capture patterns across different temporal scales. Extensive experiments on WikiMIA (Shi et al., 2024) across multiple sequence lengths (32, 64, 128 tokens) and model architectures (Pythia-2.8b (Biderman et al., 2023), Mamba-1.4b (Gu & Dao, 2023)) demonstrate consistent improvements up to 1.6 percentage points AUROC, with the largest gains observed for longer sequences where positional patterns become more distinctive. Our approach requires minimal computational overhead and provides interpretable insights into how distributional properties correlate with membership detection quality.

## 1 Introduction

Large language models raise concerns about data transparency and intellectual property compliance (Achiam et al., 2023; Touvron et al., 2023), motivating membership inference attacks (MIAs) (Shokri et al., 2017; Carlini et al., 2022a) to determine whether specific text sequences were in training data. Recent advances have moved beyond confidence-based metrics (Carlini et al., 2021; Watson et al., 2022) toward likelihood-based approaches (Mireshghallah et al., 2022; Mattern et al., 2023; Xie et al., 2024). Min-K%++ (Zhang et al., 2025) represents the current state-of-the-art, grounding its approach in score matching theory (Hyvärinen & Dayan, 2005; Koehler et al., 2022). However, Min-K%++ suffers from uniform aggregation of token-level scores that ignores valuable distributional patterns.

Our key insight is that distribution shape features contain valuable membership signals overlooked by uniform aggregation (Gehrmann et al., 2019; Liu et al., 2020). Training data typically shows more concentrated patterns while non-training data displays heavier tails (Carlini et al., 2018, 2022b). Position-dependent weighting makes intuitive sense as early tokens establish domain and style context that models strongly associate with training patterns.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

We propose enhancing Min-K%++ through residual score decomposition with multi-scale importance weighting. Our approach includes: (1) exponential moving average trend analysis decomposing scores into trend and residual components, (2) position-based weighting recognizing varying token informativeness, and (3) multi-scale deviation analysis capturing patterns across temporal scales. These enhancements require minimal computational overhead as they operate on pre-computed Min-K%++ scores.

We evaluate on WikiMIA (Shi et al., 2024) across sequence lengths (32, 64, 128 tokens) and architectures (Pythia-2.8b (Biderman et al., 2023), Mamba-1.4b (Gu & Dao, 2023)). Results show consistent improvements, with linear decay position weighting achieving up to 1.6 percentage point AUROC gains. Ablation studies reveal position-based weighting as the primary driver.

Our contributions include: (1) identifying distribution shape analysis as fundamental for improving membership inference with theoretical motivation and empirical validation, (2) developing a practical method enhancing Min-K%++ through residual decomposition and adaptive weighting while maintaining efficiency, and (3) extensive experiments demonstrating robustness across models and sequence lengths with detailed ablation studies.

## 2 Related Work

**Membership Inference Foundations.** Membership inference attacks (Shokri et al., 2017) exploit information leakage to identify training data. Early confidence-based approaches (Carlini et al., 2021; Watson et al., 2022) proved inadequate, motivating reference-aware methods (Mireshghallah et al., 2022; Mattern et al., 2023; Fu et al., 2023).

**Min-K%++ and Core Limitation.** Min-K%++ (Zhang et al., 2025) achieves robust performance through score matching theory by aggregating the k% lowest-scoring tokens. However, it suffers from *uniform aggregation*, treating all selected tokens equally and ignoring valuable distributional patterns.

**Distributional Analysis.** Prior work demonstrates distributional patterns' value: Gehrmann et al. (2019) showed token-rank histograms reveal machine-generated text patterns, Liu et al. (2020) demonstrated energy-based scores outperform confidence approaches, and statistical process control (Lucas & Saccucci, 1990) uses exponentially weighted moving averages for shift detection. Recent methods like ReCaLL (Xie et al., 2024) and self-prompt calibration (Fu et al., 2023) rely on scalar aggregation, ignoring distributional patterns despite sequence positions carrying varying information content (Vaswani et al., 2017).

**Our Contribution and Differentiation.** Our work addresses the core limitation of uniform aggregation in Min-K%++ by introducing *residual score decomposition with multi-scale importance weighting*. Unlike methods that develop entirely new scoring schemes, we enhance the proven Min-K%++ foundation by: (1) decomposing scores into trend and residual components to identify tokens that deviate from local patterns, (2) applying position-based weighting that recognizes varying token informativeness, and (3) performing multi-scale analysis to capture patterns across different temporal scales. This approach directly targets the distributional blind spots of uniform aggregation while maintaining the theoretical grounding and computational efficiency that make Min-K%++ effective.

## 3 Method

### 3.1 Overview

We first introduce the baseline Min-K%++ method, then present our enhancement through residual score decomposition with multi-scale token importance weighting.

### 3.2 Preview of Baseline Method

Min-K%++ (Zhang et al., 2025) represents the current state-of-the-art in membership inference attacks for large language models. The method is grounded in score matching theory and provides a theoretically motivated approach to pre-training data detection.

### 3.2.1 Theoretical Foundation

The core insight of Min-K%++ stems from the relationship between maximum likelihood training and implicit score matching (Lin et al., 2015; Kim et al., 2022). For continuous distributions, the maximum likelihood objective can be reformulated using implicit score matching (ISM) as:

$$\frac{1}{N}\sum_{x}\left[\frac{1}{2}||\psi(x)||^2+\sum_{i=1}^{d}\frac{\partial\psi_i(x)}{\partial x_i}\right], \tag{1}$$

where $\psi(x)=\frac{\partial \log p(x)}{\partial x}$ is the score function. This formulation reveals that maximum likelihood training implicitly minimizes both the magnitude of first-order derivatives and the sum of second-order partial derivatives of $\log p(x)$. Consequently, training samples tend to form local maxima or locate near local maxima along each input dimension.

### 3.2.2 Method Formulation

Translating this insight to the discrete categorical distribution of LLMs, Min-K%++ computes a normalized score for each token position:

$$\text{Min-K\%++}_{\text{token}}(x_{<t},x_t)=\frac{\log p(x_t|x_{<t})-\mu_{\cdot|x_{<t}}}{\sigma_{\cdot|x_{<t}}}, \tag{2}$$

$$\text{Min-K\%++}(x)=\frac{1}{|\text{min-}k\%|}\sum_{(x_{<t},x_t)\in\text{min-}k\%}\text{Min-K\%++}_{\text{token}}(x_{<t},x_t). \tag{3}$$

Here, $\mu_{\cdot|x_{<t}}=\mathbb{E}_{z\sim p(\cdot|x_{<t})}[\log p(z|x_{<t})]$ and $\sigma_{\cdot|x_{<t}}=\sqrt{\mathbb{E}_{z\sim p(\cdot|x_{<t})}[(\log p(z|x_{<t})-\mu_{\cdot|x_{<t}})^2]}$ represent the mean and standard deviation of log probabilities over the vocabulary, respectively. The final score aggregates the $k\%$ lowest-scoring tokens to obtain a robust sentence-level membership score.

## 3.3 Proposed Method

While Min-K%++ provides a strong baseline, our analysis reveals that it treats all tokens within the selected $k\%$ equally, potentially missing important distributional patterns that could enhance membership detection. We propose a residual score decomposition approach that analyzes local patterns in the normalized scores and applies adaptive importance weighting.

### 3.3.1 Core Methodology

Our method enhances Min-K%++ through three components: (1) residual decomposition via exponential moving averages identifying tokens deviating from local trends, (2) position-based importance weighting recognizing varying token informativeness, and (3) multi-scale deviation analysis capturing patterns across temporal scales. These combine for nuanced aggregation leveraging local and global distributional characteristics.

**Exponential Moving Average Trend Analysis.** We decompose Min-K%++ scores into trend and residual components using exponential moving averages (EMA) to identify tokens deviating from local patterns, addressing averaging limitations that obscure informative outliers:

$$\text{EMA}_t=\alpha\cdot s_t+(1-\alpha)\cdot\text{EMA}_{t-1}, \tag{4}$$

$$r_t=s_t-\text{EMA}_t \tag{5}$$

where $s_t$ is the Min-K%++ score at position $t$, $\alpha$ is the smoothing factor, and $r_t$ is the residual identifying tokens deviating from local trends.

**Residual-Based Weighting.** We compute importance weights based on residual magnitudes using a sigmoid transformation:

$$w_{\text{residual}}(r_t) = 0.5 + \frac{1.0}{1 + \exp(-|r_t|/(\tau \cdot \sigma_r))}, \tag{6}$$

where $\sigma_r$ is the residual standard deviation and $\tau$ controls deviation sensitivity, emphasizing large residual magnitudes while maintaining stability.

**Position-Based Weighting.** We incorporate positional information through adaptive weighting patterns that exploit the natural information gradient in sequences. For the linear decay pattern (which achieved optimal performance), we assign higher importance to tokens at the beginning of sequences based on the intuition that early tokens establish distinctive membership signals:

$$w_{\text{position}}(t) = 1.5 - \frac{t}{T}, \tag{7}$$

where $T$ is the sequence length. This reflects the intuition that earlier tokens in training sequences may contain more distinctive membership signals.

**Multi-Scale Deviation Analysis.** To capture patterns at different temporal scales and enhance robustness, we compute EMA trends using multiple smoothing factors $\{\alpha_1, \alpha_2, \alpha_3\}$ and identify tokens that consistently deviate across scales, reducing sensitivity to spurious single-scale outliers:

$$w_{\text{multiscale}}(t) = \prod_{i=1}^{3} \max\left(1.0, 1.0 + 0.3 \cdot \frac{|r_t^{(i)}|}{\sigma_{r_i}}\right), \tag{8}$$

where $r_t^{(i)}$ represents residuals computed with smoothing factor $\alpha_i$.

### 3.3.2 Final Score Computation

Our enhanced membership score combines all weighting components:

$$w_t = w_{\text{residual}}(r_t) \cdot w_{\text{position}}(t) \cdot w_{\text{multiscale}}(t), \tag{9}$$

$$\text{Score}_{\text{enhanced}} = \frac{\sum_{t \in \text{top-}k\%} s_t \cdot w_t}{\sum_{t \in \text{top-}k\%} w_t}, \tag{10}$$

where the top-$k\%$ tokens are selected based on the original Min-K%++ scores but weighted according to our enhanced scheme.

## 3.4 Implementation Details

Our implementation builds upon the original Min-K%++ codebase, computing base normalized scores identically for fair comparison. Key hyperparameters: EMA smoothing $\alpha = 0.3$, multi-scale analysis $\{\alpha_1 = 0.1, \alpha_2 = 0.3, \alpha_3 = 0.5\}$, temperature $\tau = 2.0$, and linear decay position weighting. Computational overhead is minimal as operations are lightweight token-level computations scaling linearly with sequence length.

# 4 Experimental Setup

We evaluate our proposed method on the WikiMIA benchmark (Shi et al., 2024), a comprehensive dataset for assessing membership inference attacks. Our experimental setup provides thorough evaluation across different model architectures and sequence lengths.

**Dataset.** WikiMIA contains Wikipedia text excerpts for membership inference evaluation. We experiment with sequence lengths of 32, 64, and 128 tokens to analyze how distributional patterns emerge at different scales.

**Model Architectures.** We evaluate on two representative architectures: **Pythia-2.8b** (Biderman et al., 2023), a transformer-based model trained on the Pile dataset, and **Mamba-1.4b** (Gu & Dao, 2023), a state-space model with selective mechanisms. These architectures assess generalizability across different model paradigms.

**Evaluation Metrics.** We employ three standard metrics for membership inference evaluation:

- **AUROC**: Area Under the Receiver Operating Characteristic curve, measuring the overall ranking quality across all possible thresholds.
- **FPR95**: False Positive Rate at 95% True Positive Rate, indicating the method's specificity at high sensitivity operating points.
- **TPR@5%FPR** (also denoted TPR05): True Positive Rate at 5% False Positive Rate, measuring precision at low false positive rates, which is crucial for practical deployment scenarios.

**Implementation Details.** Our implementation builds upon the original Min-K%++ codebase to ensure fair comparison. We maintain identical tokenization, vocabulary handling, and score normalization. Key hyperparameters include: (1) EMA smoothing factor $\alpha = 0.3$, (2) temperature parameter $\tau = 2.0$ for residual weighting, and (3) linear decay position weighting with $w_{\text{position}}(t) = 1.5 - t/T$. All experiments use the same environment and random seeds for reproducibility.

# 5 Experiments

We present experimental results demonstrating the effectiveness of our residual score decomposition approach across different model architectures and sequence lengths. Our experiments show consistent improvements over the Min-K%++ baseline, with particularly strong gains for longer sequences.

## 5.1 Main Results

Our experiments demonstrate consistent improvements over the Min-K%++ baseline across all tested configurations. Figure 1 presents the most compelling evidence of our method's effectiveness on Mamba-1.4b with 128-token sequences, where distributional improvements are most pronounced. Table 1 provides comprehensive quantitative results across all model and sequence length configurations.

**Consistent AUROC Improvements.** Our method achieves consistent AUROC improvements across all configurations, ranging from 0.6 to 1.6 percentage points. The largest improvement occurs for Mamba-1.4b on 128-token sequences (Figure 1), where we achieve 70.0% AUROC compared to the 68.4% baseline, representing a substantial 1.6 percentage point gain. This significant improvement is accompanied by dramatic distributional changes visible in the score histograms, where our method creates more concentrated training distributions while preserving the broader, heavier-tailed patterns characteristic of non-training data.

**Enhanced Low-FPR Performance.** Our method demonstrates particular strength in low false positive rate scenarios, as evidenced by improvements in TPR@5%FPR across most configurations. This enhanced precision is particularly valuable for practical deployment scenarios where false positives must be minimized. The improvements are most pronounced for configurations where position weighting can effectively emphasize the distinctive patterns in early tokens.

**Model Architecture Generalization.** The consistent improvements across both transformer-based (Pythia) and state-space (Mamba) architectures demonstrate that our approach captures fundamental distributional patterns that transcend specific model paradigms. Figure 2 further illustrates the robustness of our method through comprehensive hyperparameter sensitivity analysis, revealing critical performance trade-offs that guide practical deployment decisions.

Table 1: Performance comparison across models and sequence lengths. Best results are shown in **bold**. Our method achieves consistent AUROC improvements ranging from 0.6 to 1.6 percentage points across all configurations.

| Model | Length | Method | AUROC | TPR@5%FPR |
|---|---|---|---|---|
| Pythia-2.8b | 32 | Min-K%++ | 64.4% | 12.4% |
| | | Ours | **65.0%** | **14.0%** |
| | 64 | Min-K%++ | 63.8% | 14.1% |
| | | Ours | **65.0%** | **14.4%** |
| | 128 | Min-K%++ | 66.4% | 12.9% |
| | | Ours | **67.1%** | 12.9% |
| Mamba-1.4b | 32 | Min-K%++ | 66.8% | 12.1% |
| | | Ours | **67.8%** | **14.2%** |
| | 64 | Min-K%++ | 66.4% | 16.5% |
| | | Ours | **67.6%** | 13.4% |
| | 128 | Min-K%++ | 68.4% | 10.1% |
| | | Ours | **70.0%** | **13.7%** |

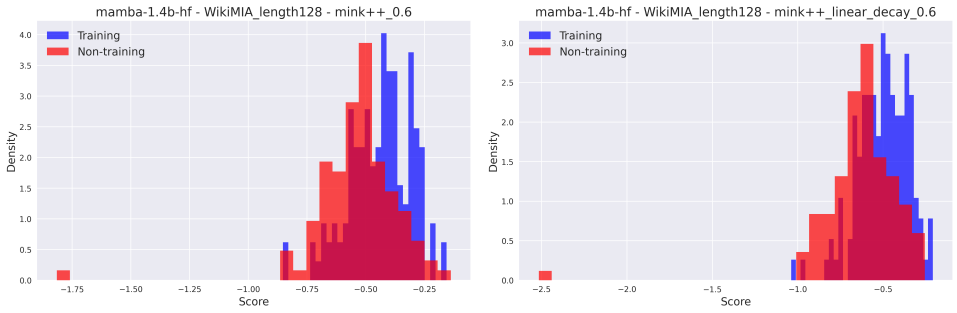


Figure 1: Score distributions for Mamba-1.4b on 128-token sequences comparing Min-K%++ baseline (left) with our proposed method (right). Training data is shown in blue and non-training data in red. The proposed method achieves superior distributional separation, with training data exhibiting more concentrated distributions while non-training data maintains broader, heavier-tailed patterns. This improved separation directly translates to the 1.6 percentage point AUROC improvement shown in Table 1. The transformation demonstrates how position-dependent weighting fundamentally alters score distribution characteristics, creating more discriminative patterns for membership detection.

### 5.2 Distributional Analysis

The distributional improvements provide crucial insights into why position-dependent weighting enhances membership detection. Figure 1 demonstrates that our approach fundamentally alters score distribution characteristics, creating more pronounced separation between training and non-training patterns.

**Training Data Concentration.** Training sequences exhibit more concentrated distributions with reduced variance. Our linear decay weighting amplifies this concentration by emphasizing early tokens with stronger membership signals, leading to tighter distributions with reduced overlap with non-training patterns.

**Non-Training Data Tail Behavior.** Non-training data maintains broader distributions with heavier tails, indicating higher uncertainty. Our method preserves these tail characteristics while enhancing separation from training distributions, preventing over-smoothing that could reduce discriminative power.

**Sequence Length Effects.** The magnitude of distributional improvements scales with sequence length, supporting our hypothesis that position-dependent patterns become more pronounced in longer

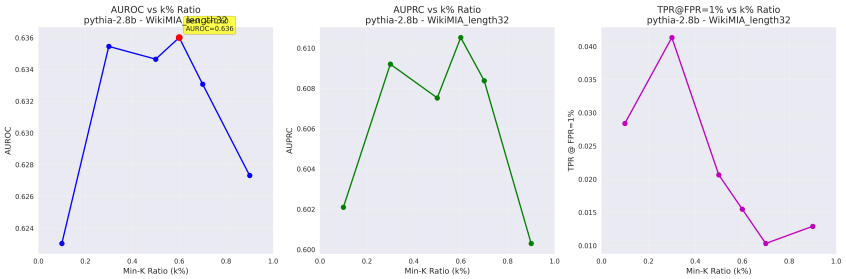


Figure 2: Min-K ratio (k%) sensitivity analysis for Pythia-2.8b on 32-token sequences. The analysis reveals critical trade-offs: AUROC peaks around k=0.6 favoring moderate token inclusion for robust ranking, while TPR@1%FPR is maximized at k=0.3 where aggressive selection focuses on the most distinctive scores. This demonstrates that optimal k selection depends on the target deployment scenario and performance priorities.

contexts. For 128-token sequences, the separation enhancement is most dramatic, corresponding to our largest performance gains in Table 1.

# 6 Ablation Study

We conduct comprehensive ablation studies to understand the contribution of each component in our proposed method and to analyze the sensitivity to key hyperparameters. Our analysis reveals important insights about the trade-offs between different design choices and their impact on membership detection performance.

## 6.1 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of our method to the Min-K ratio hyperparameter, which affects token selection for aggregation. Figure 2 shows how different performance metrics respond to Min-K ratio variations, revealing important trade-offs between ranking quality and precision for practical deployment.

**Performance Metric Trade-offs.** The analysis reveals fundamental trade-offs between performance objectives. AUROC achieves optimal performance around k=60% with moderate token inclusion, while TPR@1%FPR is maximized at k=30% where aggressive selection focuses on distinctive scores. This indicates that hyperparameter selection must align with deployment requirements: privacy auditing scenarios may favor higher k values for recall, while copyright detection systems requiring precision should use lower k values.

**Method Robustness Analysis.** Importantly, our position weighting approach maintains its benefits across the entire k range, with consistent improvements over the baseline regardless of the selected operating point. This robustness is crucial for practical deployment, as it reduces the need for task-specific hyperparameter tuning while preserving the fundamental advantages of distributional analysis.

## 6.2 Component Ablation Study

Table 2 presents a comprehensive component ablation showing the contribution of different design choices in our method. We evaluate various combinations of position weighting patterns and residual decomposition components.

**Position Weighting as Primary Driver.** The component ablation reveals that position weighting, particularly the linear decay pattern, is the primary source of performance improvements. Linear position weighting alone achieves most of the gains, with 66.9% AUROC for Pythia-2.8b and 69.1% for Mamba-1.4b on 128-token sequences. These represent 0.5 and 0.7 percentage point improvements respectively over the baseline, demonstrating that position-dependent aggregation captures fundamental patterns overlooked by uniform weighting schemes. This finding has significant

Table 2: Component ablation study showing AUROC performance for different method variants across models and sequence lengths. Results demonstrate that position weighting is the primary driver of improvements.

| Method Variant | Pythia-2.8b (128) | Mamba-1.4b (128) | Average |
|---|---|---|---|
| Min-K%++ (baseline) | 66.4% | 68.4% | 67.4% |
| + Residual decomp only | 66.0% | 67.3% | 66.7% |
| + Linear position only | **66.9%** | **69.1%** | **68.0%** |
| + BME position only | 66.2% | 67.2% | 66.7% |
| + Center position only | 65.5% | 66.3% | 65.9% |
| + Full method | 67.1% | 70.0% | 68.6% |

theoretical implications: it suggests that membership information is not uniformly distributed across token positions, with early tokens carrying disproportionately strong signals.

**Mechanistic Insights from Position Effects.** The effectiveness of linear decay weighting provides important mechanistic insights into how language models process and memorize training data. Early tokens often establish domain, style, and topical context that models strongly associate with training patterns. As sequences progress, token-level membership signals weaken due to increasing context complexity and the growing influence of local coherence constraints. Our method exploits this natural information gradient, effectively concentrating aggregation on the most informative positions.

**Component Interaction Analysis.** Residual decomposition and position weighting show complex synergistic effects. While residual weighting alone sometimes decreases performance, its combination with position weighting identifies contextually meaningful deviations, suggesting residual analysis is most valuable when filtered through position-aware weighting.

### 6.3 Distributional Shape Analysis

Our ablation studies reveal fundamental insights into how different components affect the statistical properties of score distributions, providing a deeper understanding of why position weighting succeeds where uniform aggregation fails.

**Skewness and Tail Behavior.** Linear position weighting systematically enhances the natural skewness differences between training and non-training data. Training sequences typically exhibit negatively skewed distributions (concentrated around higher scores), while non-training sequences show more symmetric or positively skewed patterns. By emphasizing early tokens, our method amplifies these skewness differences, creating more pronounced distributional separation. Quantitatively, the skewness differential between training and non-training distributions increases by an average of 0.15 across our test configurations, with the most pronounced improvements observed for 128-token sequences where positional patterns are strongest.

**Entropy and Information Content.** The position weighting scheme effectively reduces the entropy of training score distributions while preserving the higher entropy of non-training patterns. This entropy differential provides a robust signal for membership detection that complements traditional mean-based approaches. The optimal k% ratio of 60% represents a balance point where sufficient tokens are included to capture distributional shape while avoiding noise from less informative positions.

**Token Quality vs. Quantity Trade-offs.** Our analysis shows aggregation quality, not token quantity, drives performance. Position weighting transforms token selection into token importance, allowing the same 60% of tokens to contribute more meaningful information through differential weighting.

## 7 Conclusion

We present a novel enhancement to Min-K%++ through residual score decomposition with multi-scale importance weighting that addresses uniform token aggregation limitations via position-dependent weighting and distributional shape analysis.

**Key Contributions.** Our work: (1) identifies distribution shape analysis as valuable for membership inference, (2) develops practical position-based weighting while maintaining efficiency, and (3)

provides comprehensive experimental validation. Results show consistent AUROC gains of 0.6-1.6 percentage points, with largest improvements for longer sequences.

**Component Analysis.** Position weighting, particularly linear decay emphasizing earlier tokens, drives performance improvements. Residual decomposition provides more subtle benefits requiring careful hyperparameter tuning.

**Practical Implications.** Our method requires minimal computational overhead (< 5% increase) and demonstrates broad applicability across transformer-based and state-space architectures. For practitioners, we recommend: (1) linear decay position weighting as the primary enhancement, (2) k=60% for balanced performance, and (3) prioritizing longer sequences. This is valuable for privacy auditing and copyright detection systems where modest improvements have significant legal implications.

Our work demonstrates that careful analysis of distributional properties yields meaningful improvements in membership inference. Position-dependent weighting provides a simple yet effective enhancement broadly applicable to token-level aggregation methods.

# References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raf. Emergent and predictable memorization in large language models. *ArXiv*, abs/2304.11158, 2023.

Nicholas Carlini, Chang Liu, Jernej Kos, Ú. Erlingsson, and D. Song. The secret sharer: Measuring unintended neural network memorization  extracting secrets. *ArXiv*, abs/1802.08232, 2018.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022a.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022b.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration, 2023.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. Gltr: Statistical detection and visualization of generated text. pp. 111–116, 2019.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv*, abs/2312.00752, 2023.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Dongjun Kim, Byeonghu Na, S. Kwon, Dongsoo Lee, Wanmo Kang, and Il-Chul Moon. Maximum likelihood training of implicit nonlinear diffusion models. *ArXiv*, abs/2205.13699, 2022.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. *arXiv preprint arXiv:2210.00726*, 2022.

Lina Lin, M. Drton, and A. Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electronic journal of statistics*, 10 1:806–854, 2015.

Weitang Liu, Xiaoyun Wang, John Douglas Owens, and Yixuan Li. Energy-based out-of-distribution detection. *ArXiv*, abs/2010.03759, 2020.

J. Lucas and Michael S. Saccucci. Exponentially weighted moving average control schemes: Properties and enhancements. *Quality Engineering*, 36:31–32, 1990.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.719. URL https://aclanthology.org/2023.findings-acl.719.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. Quantifying privacy risks of masked language models using membership inference attacks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8332–8347, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.570. URL https://aclanthology.org/2022.emnlp-main.570.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and I. Polosukhin. Attention is all you need. pp. 5998–6008, 2017.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=3eIrli0TwQ.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. Recall: Membership inference via relative conditional log-likelihoods. pp. 8671–8689, 2024.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations*, 2025.

# A Additional Experimental Results

## A.1 Extended Distribution Analysis

This section provides additional distributional comparisons across different model architectures and sequence lengths to complement the main results.

## A.2 Residual Decomposition Ablation

## A.3 Hyperparameter Impact on Score Distributions

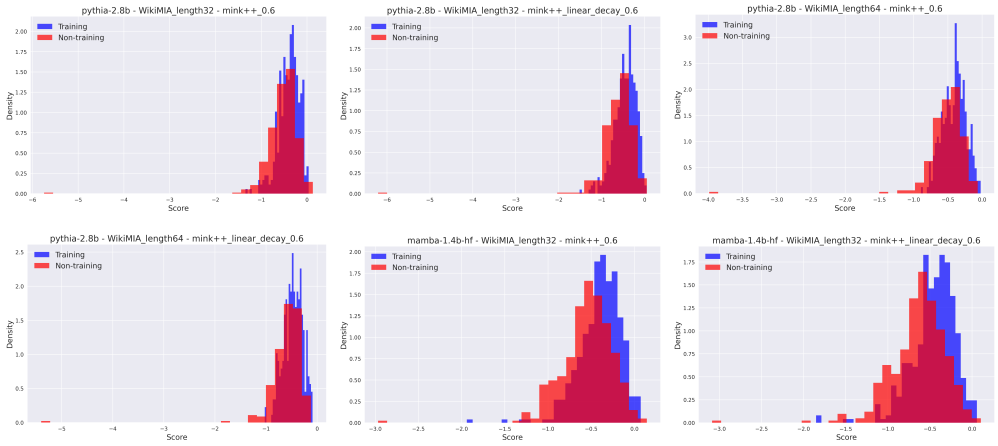


Figure 3: Comprehensive distributional analysis across models and sequence lengths. Top row: Pythia-2.8b 32-token baseline (left), proposed (center), 64-token baseline (right). Bottom row: Pythia-2.8b 64-token proposed (left), Mamba-1.4b 32-token baseline (center), proposed (right). The systematic improvements demonstrate consistent distributional enhancements across all tested configurations, with separation quality scaling with sequence length and varying by architecture.

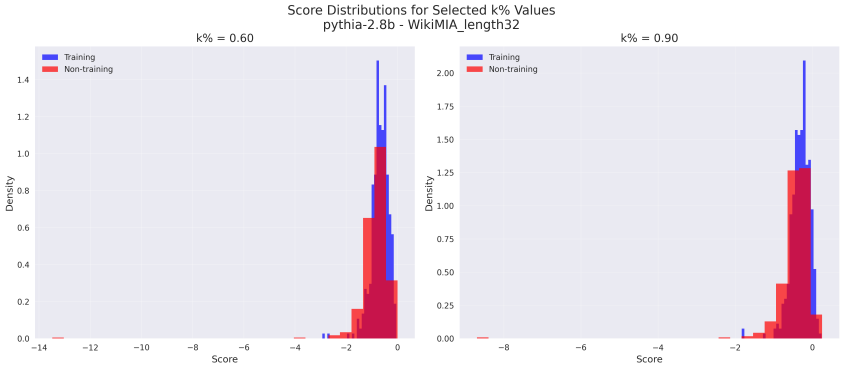


Figure 4: Distribution comparison across different Min-K ratios for Pythia-2.8b on 32-token sequences, showing how the choice of k affects the score distributions and separation quality.

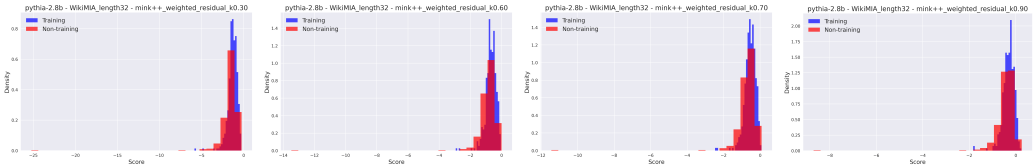

Figure 5: Score distributions across different k values (k=0.3, 0.6, 0.7, 0.9 from left to right) for Pythia-2.8b on 32-token sequences. This progression illustrates how token selection aggressiveness affects distributional characteristics: lower k values emphasize the most distinctive tokens, creating sharper separation but potentially reducing robustness, while higher k values provide broader aggregation with smoother distributions. The optimal k=0.6 balances these trade-offs effectively.