Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077179

ARTICLE

ADS: Adaptive Dataset Selection for Fine-Tuning in Anomalous Text

Xiaoyong Zhao[1], **Jiamin Wu**[2,*] **and Lei Wang**[2]

[1]School of Information Management, Beijing Information Science and Technology University, Beijing, China
[2]College of Computer Science, Beijing Information Science and Technology University, Beijing, China

*Corresponding Author: Jiamin Wu. Email: 2023020988@bistu.edu.cn

Received: 03 December 2025; Accepted: 18 March 2026; Published: 15 June 2026

ABSTRACT: With the continuous improvement of the performance of large language models, how to further enhance their ability in complex tasks has become a key issue. The task of abnormal text detection poses a challenge to the model in identifying non-standard semantics due to its semantic complexity and high-risk features. However, existing fine-tuning methods rely heavily on static data selection strategies, making it difficult to adapt to the dynamic evolution of model capabilities, resulting in low training efficiency. This article proposes ADS (Adaptive Dataset Selection), an adaptive framework for selecting data in anomaly text detection. ADS performs model-aware data selection prior to fine-tuning, adapting the initial state of pre-trained language models by selecting samples that are most informative for the target anomaly detection task. Empirical results on mainstream large language model architectures show that ADS significantly compresses data size while still outperforming existing static strategies and mainstream compression methods. When using only 1000 fine-tuning samples, ADS achieves a 92% F1 score, with an accuracy improvement of over 22% compared to the baseline, demonstrating excellent performance. This study proposes an efficient data selection mechanism from the perspective of model capability and dynamic adaptation of data, providing theoretical support and a practical path for fine-tuning large models in low-resource scenarios.

KEYWORDS: Adaptive dataset selection; anomalous text detection; fine-tuning; large language models; dynamic sample optimization; data diversity

1 Introduction

In the context of ongoing advancements in large language model (LLM) research in recent years, fine-tuning has gradually become one of the key technical pathways for enhancing a model's adaptability to downstream tasks. This method guides models to comprehend human instructions, thereby achieving stronger generalization across diverse tasks. Particularly in the complex domain of anomalous text detection, models must contend with multiple types of semantic anomalies such as fraudulent information, counterfeit content, and adversarial perturbations. How to enable models to accurately identify such texts while maintaining robustness and transferability has become a critical indicator of their practical application value. Therefore, constructing a fine-tuning dataset that can dynamically adapt to model capabilities while ensuring structural coherence and task relevance has emerged as a core issue for improving performance in such tasks.

Although recent research in LLM fine-tuning has focused on optimizing data selection strategies, mainstream approaches primarily emphasize enhancing data quality [1,2] or expanding dataset scale to improve coverage. These strategies typically rely on manually curated high-quality samples [3] or a large-scale collection of task-relevant corpora, aiming to boost model learning effectiveness and stability. However, such

static data configurations overlook a critical fact: LLMs exhibit dynamic perceptual capabilities and learning needs throughout the training process. The model's response to input data and the resulting information gain evolve non-linearly over time. This leads to a "phase mismatch" problem, where samples that are effective in early stages may become redundant or even harmful due to negative transfer after the model's capabilities improve, thus constraining overall fine-tuning efficiency and final performance.

This paper proposes Adaptive Data Selection (ADS), a model-aware data selection framework for efficient fine-tuning of large language models (LLMs) in anomalous text detection. Unlike static strategies, ADS evaluates training samples with respect to three complementary criteria—data diversity, semantic coverage, and model uncertainty—prioritizing samples that are both representative and informative for the current model state. By aligning data selection with model capability, ADS improves training efficiency and detection performance under limited data budgets. Experiments across multiple datasets and LLM backbones (e.g., LLaMA and Qwen) demonstrate that ADS consistently outperforms traditional methods using substantially fewer training samples, challenging the conventional "more data is better" assumption.

At the methodological level, ADS differs from existing adaptive or multi-factor selection approaches by introducing mechanism-level co-optimization rather than metric-level aggregation. Previous methods typically rely on static heuristics independent of training dynamics or a single adaptive signal, such as uncertainty or difficulty. In contrast, ADS decouples structural constraints and model-dependent refinement: diversity and coverage preserve the global semantic structure of the candidate pool, while model uncertainty determines sample utility relative to the model state. This design enables adaptive rebalancing between exploration and exploitation during fine-tuning, a capability not supported by prior approaches such as IFD, MoDS, or self-guided filtering.

In this work, the term adaptive refers to adapting a pre-trained language model to a downstream anomalous text detection task through model-aware data selection, rather than real-time or streaming data selection during training. Specifically, ADS evaluates candidate samples with respect to the current pre-trained model state and constructs a high-value subset prior to fine-tuning. While the selection criteria incorporate signals related to model uncertainty and semantic structure, the fine-tuning itself is conducted in an offline manner. We leave fully online or iterative re-selection across training epochs as future work.

2 Related Work

Anomaly Detection aims to identify items, events, or observations that deviate from expected patterns, representing risks such as financial fraud, industrial failures, medical anomalies, or content violations in texts [4]. Textual anomaly detection is particularly challenging due to semantic complexity and ambiguous boundaries. Traditional methods based on rules or static features struggle with semantic variation and adversarial obfuscation, while deep learning approaches, such as toxic comment detection [5] and hate speech detection [6], show improved performance but rely heavily on large annotated datasets and may lack domain generalization. Self-supervised learning methods like PANDA [7] and COCA [8] attempt to alleviate labeling scarcity but face limitations in capturing long-range semantic contradictions. Conventional anomaly detection follows a "static filtering + global optimization" paradigm [9], which is vulnerable to concept drift and adversarial tactics, leading to three main challenges: (1) lack of fine-grained semantic representation, (2) lag in dynamic adaptation, and (3) fragility of external dependencies. Large Language Models (LLMs) have shown promise in anomaly detection due to strong semantic understanding and generality, demonstrated by works like LogGPT [10], LLMAD [11], and AD-LLM [12]. However, domain-specific fine-tuning remains essential, with data selection critically influencing performance. Early approaches emphasized high-quality data, such as LIMA [3], but manual annotation is costly and risks reducing data diversity. Automated metrics

like InstructMining [13] fail to adapt dynamically to evolving training needs, while clustering methods [14] and small external models [15] have their own limitations.

Adaptive data selection methods based on model confidence, such as IFD [16], exist but often focus on single dimensions and neglect the diverse demands of anomaly detection. Comprehensive analyses [17] highlight the importance of jointly optimizing quality, diversity, and task specificity, yet most prior work treats these factors independently. An ideal data selection strategy should fulfill three criteria: (1) task adaptability to evolving long-tail anomalies; (2) endogenous motivation enabling closed-loop feedback without external reliance; and (3) multidimensional synergy optimizing diversity, coverage, and uncertainty. Prior multidimensional methods [18,19] apply these indicators statically or independently, whereas the proposed ADS models selection coupled with model evolution, using coverage and diversity as semantic scaffolds and uncertainty as dynamic control. This approach addresses LLMs' phased learning in anomaly detection and offers a new paradigm for efficient fine-tuning in low-resource settings.

3 Research Method

The research method proposed in our study consists of three phases: data preprocessing, anomaly detection, and anomaly analysis and optimization. Fig. 1 shows the overview framework of this work.

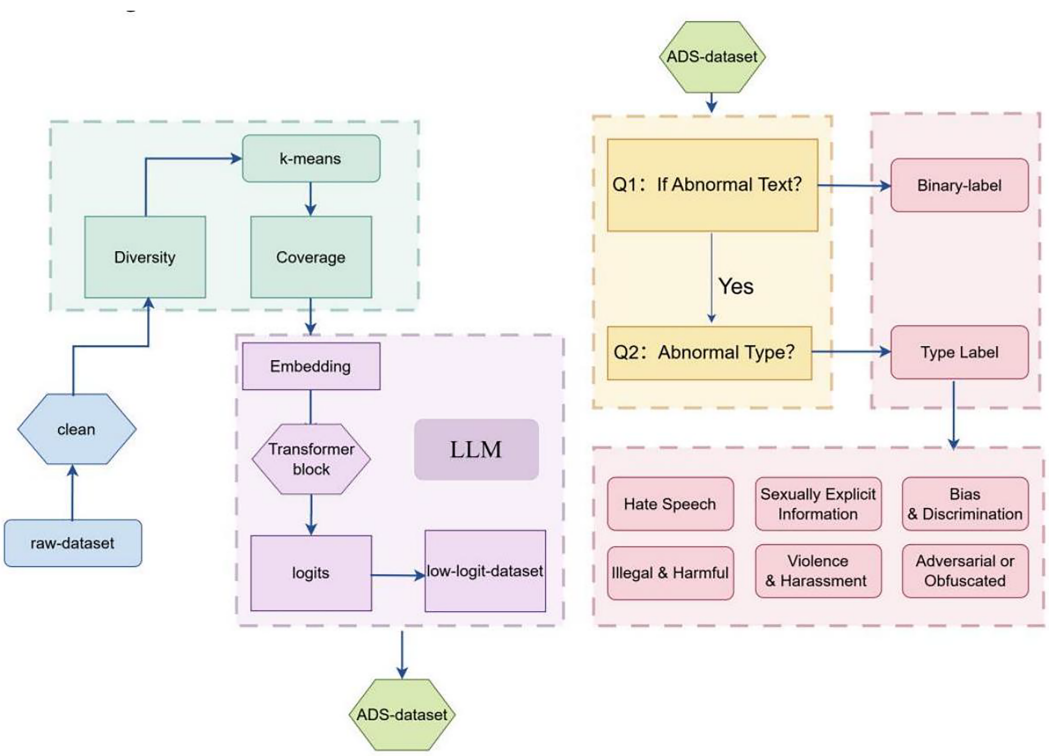

Figure 1: Overall framework of the ADS.

3.1 Definition of Anomalous Text

In the domain of textual content safety, anomalous text typically refers to content that significantly deviates from mainstream corpora in terms of language style, semantic structure, or emotional tone, and may pose threats to platform security, user psychology, or social ethics. Such text is not merely semantically deviant; it often embodies complex characteristics such as informational attacks, value conflicts, or deliberate

evasion of detection mechanisms [20]. To systematically define "anomalous text," this paper draws on Google's content safety policies [21] and existing content classification frameworks, focusing on scenarios related to online information security. We propose six representative categories of anomalous text, as outlined in Table 1:

Table 1: Categories of anomalous text.

No.	Type	Description
1	Hate Speech	Text that targets specific individuals or groups with insulting, discriminatory, or inciteful language based on characteristics such as race, religion, gender, or sexual orientation. These messages are highly inflammatory and socially harmful, potentially inciting public conflict.
2	Sexually Explicit Information	Text that directly or implicitly depicts sexual behavior, anatomy, or sexually suggestive content. Such content is particularly sensitive, poses challenges for moderation, and may negatively affect underage users or violate the platform policies.
3	Bias and Discrimination	Text based on stereotypes or social prejudice, offering one-sided evaluations of specific groups—e.g., gender bias, ageism, or occupational stigma. These messages are often subtle but pose real threats to social fairness and equity.
4	Illegal and Harmful Content	Text that involves incitement to crime, rumor spreading, privacy breaches, terrorist propaganda, drug promotion, or other illegal content. Such texts are directly harmful and pose high risks.
5	Violence and Harassment	Text that uses threatening language, describes violence, or engages in personal attacks, aiming to instill fear, cause harm, or apply sustained psychological pressure. Common on social media or malicious comments.
6	Adversarial or Obfuscated Content	Text that deliberately evades content filters through tactics such as homophones, word splitting, or semantic misdirection. These texts are adversarial in nature and challenge the robustness of detection models.

Unlike traditional anomaly detection, identifying anomalous text must account for contextual dependency, semantic diversity, and long-tail distributions, placing higher demands on large language models (LLMs) in terms of semantic comprehension, implicit intent recognition, and moral judgment. This complexity underscores the crucial role of high-quality, well-structured fine-tuning datasets in unleashing the full potential of LLMs.

3.2 Fine-Tuning Data Indicators

Fine-tuning is essential for improving large language models' (LLMs) generalization in specialized tasks like anomalous text detection, where adaptive dataset design is critical. This paper proposes a data selection framework based on three key dimensions: (1) Diversity—ensuring varied anomaly types and linguistic styles to prevent overfitting; (2) Coverage—representing the corpus space, especially long-tail, high-risk samples to boost robustness; and (3) Model Uncertainty—quantifying ambiguous or difficult inputs via confidence measures to refine decision boundaries. Unlike existing static methods, the proposed Adaptive Data Selection (ADS) method dynamically couples sample selection with the evolving model state, enhancing training efficiency and performance under limited data conditions.

3.3 Data Collection and Filtering

To fully unleash the potential of large language models (LLMs) in anomalous text detection tasks, the rational design and filtering of fine-tuning datasets is a key step toward enhancing model performance. Given LLMs' sensitivity to input data and the inherent diversity and complexity of anomalous texts, this paper proposes the ADS (Adaptive Data Selection) method. ADS integrates two core dimensions—data diversity and coverage—and model prediction uncertainty to adaptively filter and optimize fine-tuning samples, thereby improving model generalization and robustness.

3.3.1 Diversity

Diversity refers to the heterogeneity of samples across dimensions such as semantics, structure, and task intent. It is crucial for constructing robust fine-tuning datasets, enabling large language models to thoroughly explore the boundaries of the input space, thereby enhancing their generalization ability to variable input patterns [22,23]. In the ADS method, to ensure broad coverage of the corpus space during sample selection, we apply the K-means clustering algorithm on the original dataset, dividing the samples into μ_k semantic clusters $\mathcal{C} = C_1, C_2, C_K$. The clustering centers are optimized by minimizing the sum of squared intra-cluster distances:

$$\min_{\mu_1,,\mu_K} \sum_{k=1}^{K} \sum_{\mathbf{v}_j \in C_k} \left\| \mathbf{v}_j - \mu_k \right\|_2^2 \tag{1}$$

This process ensures high semantic coherence within each cluster and sufficient semantic divergence across clusters, thus enabling structured partitioning and representation of the corpus space. This mechanism effectively avoids the phenomenon of training samples being overly concentrated on "templated" anomaly types, thereby enhancing the model's ability to handle complex and compositional anomaly structures.

3.3.2 Coverage

In ADS, semantic coverage measures how well a selected subset preserves the global semantic distribution of the original corpus in the embedding space. Unlike diversity, which focuses on pairwise dissimilarity, coverage emphasizes distributional representativeness, aiming to prevent fine-tuning from overfitting to high-frequency patterns while neglecting rare but informative semantic structures [24]. Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ denote the full corpus and $\mathcal{S} \subset \mathcal{D}$ the selected subset. Each sample is mapped to a semantic embedding $\mathbf{v}_i$. Semantic coverage is approximated by jointly considering global centrality and local density within semantic clusters:

$$\mathcal{C}(\mathcal{S}) = \sum_{x \in \mathcal{S}} \left(\alpha \, \text{Centrality}(x) + (1 - \alpha) \, \text{Density}(x) \right) \tag{2}$$

where $\alpha \in [0,1]$ balances the two components.

To operationalize this objective, the corpus is partitioned into K clusters $\{C_k\}_{k=1}^{K}$. Within each cluster, the top-n representative samples are selected to form the coverage-driven subset. For a sample $x_j \in C_k$, the centrality score is defined as:

$$\text{Centrality}(x_j) = -\|\mathbf{v}_j - \boldsymbol{\mu}_k\|_2 \tag{3}$$

where $\boldsymbol{\mu}_k$ is the centroid of cluster C_k, favoring samples close to the dominant semantic region.

Local density captures neighborhood concentration within the cluster:

$$\text{Density}(x_j) = \sum_{\mathbf{v}_l \in C_k} \exp\left(-\frac{\|\mathbf{v}_j - \mathbf{v}_l\|_2^2}{2\sigma^2}\right) \tag{4}$$

where σ controls the neighborhood scale. Samples are ranked by a weighted combination of centrality and density, and the top-n samples per cluster are selected.

By explicitly modeling semantic coverage in this manner, ADS ensures that the fine-tuning data remain distributionally aligned with the original corpus, mitigating bias caused by skewed or redundant samples. The resulting coverage-driven subset further serves as a high-quality candidate pool for subsequent uncertainty-based filtering, following a "coverage first, refinement later" selection paradigm.

3.3.3 Uncertainty

After obtaining a candidate sample set with sufficient structural diversity and semantic coverage, ADS further incorporates a model-aware selection mechanism to maximize training utility under the current model parameters. Unlike traditional static sampling, ADS dynamically evaluates each sample's potential to provide information gain for the model by prioritizing samples that induce a significant increase in model prediction uncertainty [25]. The uncertainty metric used here is the entropy $\mathcal{H}(x)$, which quantifies the width of the model's prediction confidence interval for a given sample. It is defined as:

$$H(x) = -\sum_{i=1}^{C} p_i(x) \log p_i(x) \tag{5}$$

where $-(p_i(x))$ is the model's predicted probability for class i, and C is the total number of classes. To ensure comparability across different samples, we introduce the concept of normalized entropy:

$$H_{\text{normalized}}(x) = \frac{H(x)}{\log C} \tag{6}$$

Furthermore, we combine model uncertainty with actual prediction outcomes to define the final training set $\mathcal{D}_{\text{ADS}}$ as:

$$\mathcal{D}_{\text{ADS}} = \{x_i \mid \text{Incorrect}(x_i) \vee \mathcal{H}(x_i) > \beta\} \tag{7}$$

The indicator incorrect(x) is computed on the labeled training pool during data selection, and is used solely to identify samples that the current model fails to classify correctly. While this setting is offline, it simulates a realistic fine-tuning scenario where labeled data is available. Extending ADS to a fully online or streaming selection is left for future work.

The dominant computational cost of ADS (Algorithm 1) arises from embedding computation $\mathcal{O}(|\mathcal{D}| \cdot d)$ and K-means clustering $\mathcal{O}(K \cdot |\mathcal{D}| \cdot d)$, both executed once before fine-tuning. The entropy-based filtering

stage scales linearly with the size of the candidate pool. In practice, ADS introduces negligible overhead compared with LLM fine-tuning. The number of clusters K controls semantic granularity, and per-cluster sampling ensures balanced coverage. The density bandwidth is set proportional to the average intra-cluster distance, enabling adaptation to the corpus scale, while the entropy selection ratio β determines the proportion of uncertainty-driven samples and is decayed during training to balance exploration and convergence. These design choices ensure reproducibility and robustness without excessive tuning complexity (see Appendix A for detailed hyperparameter settings and dataset statistics).

Algorithm 1: Adaptive Dataset Selection (ADS)

Require: Dataset $\mathcal{D}$, model M, number of clusters K, samples per cluster n
Ensure: Selected training set $\mathcal{S}$
1: Encode all samples in $\mathcal{D}$ using a pretrained encoder
2: Perform K-means clustering to obtain clusters $\{\mathcal{C}_k\}_{k=1}^{K}$
3: **for** each cluster $\mathcal{C}_k$ **do**
4: Select n representative samples based on centrality and density
5: **end for**
6: Construct candidate pool $\mathcal{P} = \bigcup_{k=1}^{K} \mathcal{C}_k$
7: **for** each sample $x \in \mathcal{P}$ **do**
8: Compute prediction entropy $H(x)$ using the current model M
9: **end for**
10: Select the top β% samples with the highest normalized entropy
11: Fine-tune model M on the selected set $\mathcal{S}$

By integrating diversity, coverage, and uncertainty, ADS constructs a compact yet informative fine-tuning dataset $\mathcal{D}_{\text{ADS}}$ characterized by semantic breadth, category coverage, and informational density. This dataset is presented to LLMs (e.g., LLaMA-3-8B-Instruct and Qwen2.5-7B) using standardized prompt templates for binary anomaly classification. From a unified perspective, diversity and coverage define a structural feasible region preserving global semantic properties, while uncertainty acts as a model-dependent utility function for ranking samples within this region. Experimental results show that fine-tuning with only 1000 ADS-selected samples achieves an F1 score of 92% on the test set, significantly outperforming traditional static data selection methods under low-resource conditions.

3.4 Fine-Tuning Architecture and Training Process

The ADS framework operates as a data selection mechanism for supervised fine-tuning rather than an inference-time prompting strategy. After ADS identifies a high-value subset of training samples, these samples are used to fine-tune large language models via parameter-efficient adaptation.

As illustrated in Fig. 2, the backbone LLM parameters are kept frozen during training, while Low-Rank Adaptation (LoRA) modules are inserted into the attention projection layers. Specifically, LoRA adapters are applied to the query and value projection matrices of each transformer layer, enabling task-specific adaptation with a small number of trainable parameters. ADS-selected samples are formatted using a unified prompt template and fed into the model during supervised fine-tuning. Model parameters are updated by minimizing the cross-entropy loss between predicted labels and ground-truth anomaly annotations. This design decouples data selection from model architecture while ensuring that performance gains arise from

parameter-level optimization rather than prompt engineering. By integrating ADS with LoRA-based fine-tuning, the framework achieves efficient adaptation under low-resource settings, balancing training cost and detection performance.

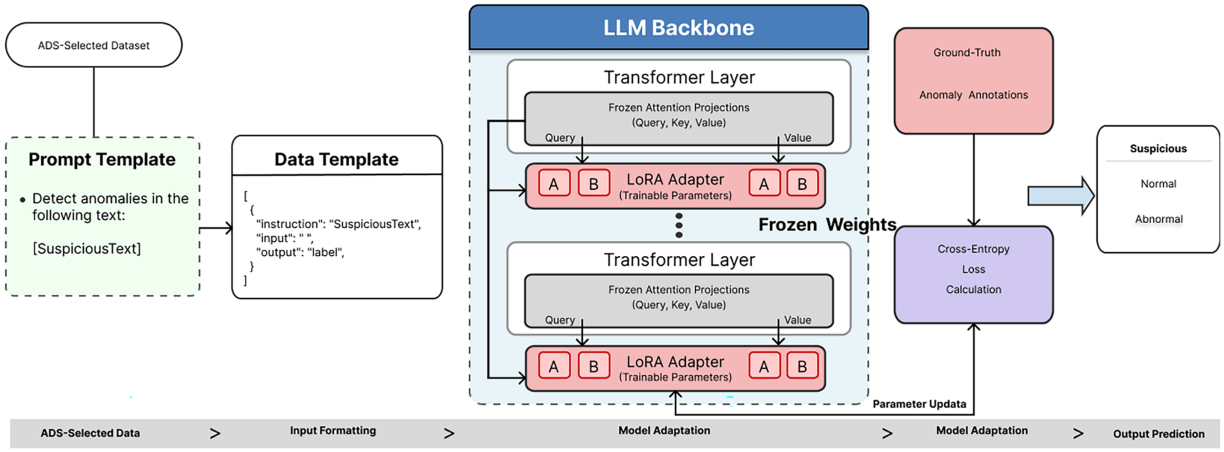

Figure 2: Fine-tuning architecture with ADS.

4 Experiments

4.1 Dataset and Baseline Models

Although several publicly available benchmark datasets exist for content safety, they often fall short of meeting the needs of anomalous text detection in terms of semantic complexity, multimodality, and linguistic diversity. To address this, we construct a new multi-source hybrid anomalous text dataset as the foundational corpus for instruction fine-tuning. The dataset is a hybrid of publicly available sources and curated subsets. All sources comply with their original licenses and are used for research purposes only. Potential biases toward commonly reported anomaly types remain, and results should be interpreted accordingly. This dataset integrates anomalous texts from both Chinese and English sources, covering a wide range of categories, including but not limited to: Sexually Explicit Information, Bias and Discrimination, Illegal and Harmful Content, Violence and Harassment, Adversarial or Obfuscated Content. Each data instance is organized as a structured triplet consisting of: Text (the original input), Label (binary classification $\in \{0, 1\}$), Type (the specific anomaly category).

Table 2 reports the distribution of anomalous text categories in the constructed dataset. Each instance is annotated with a binary anomaly label as well as a fine-grained anomaly type. The dataset integrates samples from multiple public sources and covers both Chinese and English texts, enabling evaluation under multilingual and semantically diverse settings. As shown in Table 2, the dataset exhibits a balanced yet realistic distribution across different anomaly categories, including both high-frequency and long-tail types. Approximately 34.1% of the samples are in English and 65.9% in Chinese, with an average text length of 109.39 tokens. This diversity reflects real-world content safety scenarios, where anomalous texts vary significantly in form, intent, and linguistic structure. The dataset is split into training, validation, and test sets with a ratio of 8:1:1. All splits are stratified by anomaly category and language to preserve distributional balance. Near-duplicate samples across splits are removed using embedding-based similarity filtering to prevent data leakage.

Table 2: Statistical overview of the multi-source hybrid anomalous text dataset.

Type	Number of Samples
Adversarial or Obfuscated Content	25,574
Bias and Discrimination	10,703
Hate Speech	17,430
Illegal and Harmful Content	19,983
Sexually Explicit Information	7137
Violence and Harassment	25,726
Total	106,553

Compared to existing datasets with single-label or monolingual structures, our dataset provides higher semantic complexity and broader language coverage, thereby offering a more challenging and realistic environment for LLM training and evaluation. For the model setup, we selected the following three pre-trained language models as experimental baselines:

- LLaMA-3-8B-Instruct: An open-source multilingual instruction-tuned model, representative of current mainstream Instruct-tuning architectures.
- Qwen2.5-7B: A bilingual (Chinese-English) model with strong capabilities in Chinese language understanding and instruction following.
- Qwen2ForSequenceClassification: A customized version of Qwen with its architecture adapted to a sequence classification head.

All models were guided via standardized prompt templates and configured to output binary classification results (anomalous/non-anomalous). Evaluation metrics include Precision (P), Recall (R), Macro-F1 score, and Accuracy.

4.2 Overall Binary Classification Results and Analysis

The overall binary classification results are presented in Table 2. All models trained on 1000 samples selected via the ADS method achieved significantly better performance compared to both baseline models and models trained on randomly selected data of the same size. This confirms that the ADS strategy enhances model generalization while reducing data redundancy. We emphasize that ADS is designed for parameter-level fine-tuning rather than prompt-only inference. All reported results (except prompt-only baselines) are obtained by updating model parameters via supervised fine-tuning on ADS-selected datasets. Prompt templates are used solely to standardize input-output formatting and do not constitute the primary source of performance gains.

Table 3 clearly demonstrates the effectiveness of the ADS data selection method: Across all model architectures, using only 1000 ADS-selected samples yields better results than training from scratch or using 1000 random samples. Among all combinations, the LLaMA-3-8B-Instruct model trained with ADS data achieved the best overall performance. These results validate that ADS can dramatically reduce training costs while improving performance by selecting high-value, diverse, and informative training samples. This approach is especially valuable in low-resource or time-constrained deployment scenarios.

Table 3: Overall binary classification performance.

Method	Accuracy	F1	Precision (P)	Recall (R)
Llama-3-8B-Instruct + prompt	0.58	0.53	0.53	0.53
Llama-3-8B-Instruct + ADSdata	0.93	0.92	0.90	0.93
Llama-3-8B-Instruct + randomdata	0.83	0.80	0.79	0.81
Qwen2.5-7B + prompt	0.78	0.75	0.72	0.78
Qwen2.5-7B + ADSdata	0.92	0.88	0.89	0.91
Qwen2.5-7B + randomdata	0.76	0.71	0.70	0.73
Qwen2ForSequenceClassification + ADSdata	0.90	0.87	0.87	0.89
Qwen2ForSequenceClassification + randomdata	0.81	0.69	0.68	0.70

4.3 Ablation Study

To validate the individual contributions of each component in the ADS method, we conducted a series of ablation experiments. These included: the full ADS method (diversity and coverage-based clustering + model uncertainty-based filtering), using only diversity and coverage filtering, using only model uncertainty filtering, and random sampling as a baseline. We also evaluated combinations of two criteria: diversity + coverage, diversity + entropy, and coverage + entropy. The goal was to analyze how each selection strategy and its combinations impact anomaly detection performance. In addition, we examined the influence of different training data volumes (500, 1000, 5000, and the full dataset) on the effectiveness of ADS.

4.3.1 Results and Analysis

As shown in Table 4, the complete ADS strategy outperforms all other ablation variants across every evaluation metric. This confirms that simultaneously considering diversity, coverage, and model uncertainty leads to the best performance. While individual selection strategies offer some improvement over baseline, none match the effectiveness of the full ADS method. Interestingly, the random sampling group generally outperforms the single-factor strategies, suggesting that relying solely on one selection dimension can result in biased or incomplete data, whereas randomly sampled data, although more balanced, still lack the task alignment and efficiency of targeted filtering. Among the combination strategies, the coverage + entropy group performs best, achieving significant improvements in Precision, Recall, Macro F1, and Accuracy. This highlights a strong synergy between data representativeness and uncertainty-driven learning. Specifically, coverage ensures a balanced distribution across the data space, while entropy directs the model's attention to ambiguous decision boundaries, together enabling finer-grained anomaly detection. These findings indicate that coverage and uncertainty are the primary drivers of performance gains in the ADS tri-strategy setup.

Further analysis shows that the three filtering dimensions contribute in distinct ways: diversity and coverage improve generalization and breadth of anomaly types covered, whereas entropy strengthens sensitivity to ambiguous or uncertain cases. Their combination supports both efficient data space exploration and fine-grained boundary modeling, maximizing data utility.

In conclusion, these experiments validate the necessity and effectiveness of ADS's "three-dimensional co-optimization" design. Compared to any two-factor combinations, integrating diversity, coverage, and uncertainty offers better alignment with both model training objectives and task-specific challenges. This is especially critical for anomaly detection, which often involves semantic ambiguity, sparse distributions, and concept drift. Building a data selection mechanism that is both task-aware and model-guided is essential for efficient LLM fine-tuning and robust performance enhancement.

Table 4: Ablation study results.

Method	Accuracy	F1	Precision (P)	Recall (R)
ADS (Full Strategy)	0.90	0.93	0.92	0.93
Diversity Only	0.77	0.82	0.78	0.82
Coverage Only	0.76	0.82	0.78	0.82
Entropy Only	0.77	0.82	0.79	0.82
Random Sampling	0.79	0.83	0.80	0.83
Diversity + Coverage	0.80	0.85	0.82	0.85
Diversity + Entropy	0.83	0.88	0.85	0.87
Coverage + Entropy	0.87	0.91	0.88	0.90

4.3.2 Impact of Training Data Volume on Performance

This section investigates how different training data volumes affect the performance of the ADS method. Under consistent LLM conditions, we trained the model using 500, 1000, and 5000 samples, respectively, and observed changes in performance.

From the results, we can see that increasing the training set size from 500 to 1000 leads to a significant performance boost. However, further increasing the data size to 5000 results in only marginal improvements, indicating that the ADS method is capable of achieving effective training even with relatively small, high-quality datasets, without the need for large-scale data collections.

An interesting observation in Table 5 is that fine-tuning with the full dataset of approximately 80,000 samples yields worse performance than using a carefully selected subset of only 1000 samples. While this phenomenon may appear counterintuitive, it is consistent with known challenges in anomaly detection and instruction tuning. First, large-scale anomalous text datasets constructed from heterogeneous sources inevitably contain semantic redundancy, where many samples convey near-duplicate patterns with a limited incremental learning signal. Such redundancy can bias the optimization process toward dominant or frequent anomaly patterns, reducing sensitivity to informative boundary cases. Second, label noise and annotation inconsistency accumulate with dataset scale, particularly in content-safety scenarios involving subjective categories. When used indiscriminately for fine-tuning, noisy labels can distort the decision boundary and degrade generalization, especially under parameter-efficient adaptation such as LoRA. From an optimization perspective, fine-tuning on a large but noisy and redundant dataset may amplify gradient variance and induce overfitting to spurious correlations. In contrast, ADS explicitly prioritizes samples that are diverse, semantically representative, and uncertain under the current model state, effectively filtering out low-information or misleading instances. These results support the emerging "less-is-more" paradigm in efficient LLM adaptation: for specialized detection tasks, a small set of high-quality, high-information samples can yield better generalization than a much larger but uncurated corpus.

4.4 Summary

This paper introduces a novel Adaptive Data Selection (ADS) method for instruction fine-tuning in anomalous text detection. By jointly leveraging data diversity, semantic coverage, and model prediction uncertainty, ADS dynamically identifies the most informative instruction samples from large-scale datasets. This enables models to achieve superior detection performance while using significantly fewer training samples. Our experiments demonstrate the effectiveness of ADS: data selected by ADS consistently outperforms random sampling in enhancing LLM performance on anomalous text detection tasks. Moreover, combining

diversity, coverage, and uncertainty yields better results than relying on any single criterion alone. ADS also reduces redundant data, achieving strong performance even with smaller datasets.

Table 5: Performance under different training data volumes.

Training Data Volume	Precision (P)	Recall (R)	Macro F1	Accuracy
500	0.79	0.83	0.80	0.84
1000	0.90	0.93	0.92	0.93
5000	0.88	0.92	0.89	0.91
80,000 (full)	0.77	0.82	0.78	0.82

For future work, we plan to: Integrate self-supervised learning techniques to further improve selection precision. Evaluate the scalability and adaptability of ADS across LLMs of varying sizes. Explore broader applications such as scam message detection, misinformation identification, and other real-world anomaly detection scenarios. In this work, adaptivity refers to model-aware data valuation rather than epoch-level online re-sampling. ADS prioritizes samples with high uncertainty relative to the pretrained or partially adapted model, aligning dataset selection with the model's current capabilities. While the current study focuses on a single-round instantiation for clarity and stability, the framework naturally extends to multi-round or streaming settings, which we leave for future exploration.

In conclusion, ADS provides an efficient data selection framework for fine-tuning large language models, significantly reducing training costs while maintaining strong performance. Future efforts will aim to broaden ADS's applicability through self-supervised methods, evaluate its generalization on larger LLMs, and apply it across diverse anomalous text detection tasks.

Acknowledgement: I sincerely thank all individuals and institutions that supported this research beyond what is noted in the Author Contributions and Funding Statement. Gratitude goes to the administrative staff of Beijing Information Science and Technology University for their assistance in logistics, scheduling, and documentation, which ensured smooth project progression. I also appreciate the technical team at the ISI lab for their expertise in equipment operation and technical troubleshooting, which guaranteed data reliability.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Xiaoyong Zhao and Jiamin Wu; methodology, Xiaoyong Zhao and Jiamin Wu; software, Xiaoyong Zhao and Jiamin Wu; validation, Xiaoyong Zhao and Lei Wang; formal analysis, Xiaoyong Zhao and Jiamin Wu; investigation, Xiaoyong Zhao and Jiamin Wu; resources, Jiamin Wu; data curation, Jiamin Wu; writing—original draft preparation, Jiamin Wu; writing—review and editing, Jiamin Wu; visualization, Xiaoyong Zhao; supervision, Xiaoyong Zhao; project administration, Xiaoyong Zhao; funding acquisition, Xiaoyong Zhao. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data is openly available in a public repository. All data during this study are publicly available at https://www.kaggle.com/datasets/min0619/anomalous-text. Reasonable requests for additional data can be directed to the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Hyperparameters and Dataset Details

Table A1 summarizes the key hyperparameters and design choices of ADS. All parameters are empirically fixed across experiments unless otherwise stated. ADS is instantiated in a single-round offline setting, where data selection is performed once before fine-tuning to isolate the effect of adaptive data valuation. Implementation Details of ADS. All samples are encoded using the Qwen2.5 model for semantic clustering and coverage estimation. The embedding model is kept fixed to decouple representation learning from data selection. The number of clusters K scales with $\sqrt{|\mathcal{D}|}$, and a fixed top-n samples are selected per cluster to ensure balanced semantic coverage. The candidate pool size is maintained at approximately 2–3× the final training budget.

Table A1: Hyperparameter settings of ADS.

Parameter	Meaning	Value/Strategy
K	Semantic clusters	$\sqrt{\|\mathcal{D}\|}$ (proportional)
n	Samples per cluster	Fixed ratio, yielding candidate pool ≈ 2–3× budget
σ	Density bandwidth	Mean intra-cluster distance
β	Entropy ratio	Linearly decayed from 30% to 10%
Update	Re-selection	Single offline round

All experiments use identical fine-tuning configurations to ensure fair comparison. Reported results are averaged over three random seeds unless otherwise specified. Table A2 shows the dataset details, and Table A3 lists the fine-tuning settings.

Table A2: Dataset details.

Type	Lang.	Label	Num.
Adversarial/Obfuscated	Zh	0/1	18,005/1995
	En	0/1	4827/747
Bias and Discrimination	En	0/1	9507/1196
Hate Speech	Zh	0/1	9946/7484
Illegal/Harmful	En	0/1	19,277/706
Sexually Explicit	Zh	0/1	4112/3025
Violence/Harassment	Zh	0/1	13,003/12,723
Total			106,553

Table A3: Fine-tuning configuration.

Item	Setting
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	8
Epochs	3
Warmup	5%
Weight decay	0.01
Max length	512
PEFT	LoRA ($r = 8$, $\alpha = 16$)
Seeds	$\{42, 43, 44\}$
Hardware	RTX 5090 32 GB

References

1. Agarwal S, Almeida D, Askell A, Christiano P, Hilton J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–44. doi:10.52202/068431-2011.
2. Liu Z, Ke R, Liu Y, Jiang F, Li H. Take the essence and discard the dross: a rethinking on data selection for fine-tuning large language models. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2025 Apr 29–May 4; Albuquerque, New Mexico.
3. Zhou C, Liu P, Xu P, Iyer S, Sun J, Mao YN, et al. Lima: less is more for alignment. Adv Neural Inf Process Syst. 2023;36:55006–21.
4. Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Comput Surv. 2009;41(3):1–58. doi:10.1145/1541880.1541882.
5. Risch J, Krestel R. Toxic comment detection in online discussions. In: Deep learning-based approaches for sentiment analysis. Singapore: Springer; 2020. p. 85–109.
6. Caselli T, Basile V, Mitrović J, Granitzer M. HateBERT: retraining BERT for abusive language detection in English. In: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021); 2021 Aug 6; Stroudsburg, PA, USA: Association for Computational Linguistics. p. 17–25.
7. Jiang Y, Cao Y, Shen W. Prototypical learning guided context-aware segmentation network for few-shot anomaly detection. IEEE Trans Neural Netw Learn Syst. 2025;36(7):12016–26. doi:10.1109/tnnls.2024.3463495.
8. Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y. Coca: contrastive captioners are image-text foundation models. arXiv:2205.01917. 2022.
9. Pang G, Shen C, Cao L, Van Den Hengel A. Deep learning for anomaly detection: a review. ACM Comput Surv. 2022;54(2):1–38. doi:10.1145/3439950.
10. Han X, Yuan S, Trabelsi M. LogGPT: log anomaly detection via GPT. In: Proceedings of the 2023 IEEE International Conference on Big Data (BigData); 2023 Dec 15–18; Sorrento, Italy. p. 1117–22.
11. Liu J, Zhang C, Qian J, Ma MH, Qin S, Bansal C, et al. Large language models for time series anomaly detection and interpretation. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2025 Aug 3–7; Toronto, ON, Canada. p. 2145–56.
12. Yang T, Nian Y, Li S, Xu R, Li Y, Li J, et al. AD-LLM: benchmarking large language models for anomaly detection. arXiv:2412.11142. 2024.
13. Pan X, Huang L, Kang L, Liu ZC, Lu Y, Cheng SB. G-dig: towards gradient-based diverse and high-quality instruction data selection for machine translation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2024 Aug 11–16; Bangkok, Thailand. p. 15395–406.
14. Chen H, Zhang Y, Zhang Q, Yang HT, Hu XM, Ma XT, et al. Maybe only 0.5% data is needed: a preliminary exploration of low training data instruction tuning. arXiv:2305.09246. 2023.

15. Du Q, Zong C, Zhang J. Mods: model-oriented data selection for instruction tuning. arXiv:2311.15653. 2023.
16. Chen L, Li S, Yan J, Wang H, Gunaratna K, Yadav V, et al. AlpaGasus: training a better alpaca with fewer data. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR); 2024 May 7; Vienna, Austria.
17. Liu W, Zeng W, He K, Jiang Y, He J. What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning. In: Proceedings of the 12th International Conference on Learning Representations (ICLR); 2024 May 7–11; Vienna, Austria.
18. Yu S, Chen L, Ahmadian S, Fadaee M. Diversify and conquer: diversity-centric data selection with iterative refinement. arXiv:2409.11378. 2024.
19. Yang X, Nie S, Liu L, Gururangan S, Karn U, Hou R, et al. Diversity-driven data selection for language model tuning through sparse autoencoder. arXiv:2502.14050. 2025.
20. Hodge V, Austin J. A survey of outlier detection methodologies. Artif Intell Rev. 2004;22(2):85–126. doi:10.1023/b:aire.0000045502.10941.a9.
21. Zeng W, Liu Y, Mullins R, Peran L, Fernandez J, Harkous H, et al. ShieldGemma: generative AI content moderation based on Gemma. arXiv:2407.21772. 2024.
22. Ash JT, Zhang C, Krishnamurthy A, Langford J, Agarwal A. Deep batch active learning by diverse, uncertain gradient lower bounds. In: Proceedings of the International Conference on Learning Representations (ICLR); 2020 Apr 26–30; Addis Ababa, Ethiopia.
23. Bukharin A, Li S, Wang Z, Yang JF, Yin B, Li X, et al. Data diversity matters for robust instruction tuning. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; 2024 Nov 12–16; Miami, FL, USA. p. 3411–25. doi:10.18653/v1/2024.findings-emnlp.195.
24. Lin H, Bilmes J. A class of submodular functions for document summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics; 2011 Jun 19–24; Portland, OR, USA. p. 510–20.
25. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630(8017):625–30. doi:10.1038/s41586-024-07421-0.