Calibrating Transformer Attention via Task-Space Sensitivity Feedback

Anonymous ACL submission

Abstract

Transformer-based pre-trained language models (PLMs) excel in text classification but suffer from attention dilution and attention sink effects, forcing models to over-focus on task-irrelevant tokens. Existing attention supervision methods rely on costly token-level human annotations or static heuristics, which fail to scale or capture context-dependent token importance. To address this, we propose AttCal, a self-supervised, annotation-free attention calibration framework via task-space sensitivity feedback. AttCal treats attention distributions as stochastic policies to guide token deletion perturbations, infers context-dependent token importance from semantic shifts, and optimizes attention parameters via policy gradient. Critically, AttCal freezes the PLM backbone and updates only attention projection layers to ensure training efficiency. Benchmarks across five datasets demonstrate that AttCal significantly outperforms vanilla PLMs—yielding up to a 15.8% accuracy lift for Llama3-8B—and surpasses SOTA supervised baselines in both classification and interpretability.

1 Introduction

Transformer-based pre-trained language models (PLMs) have achieved remarkable success across diverse natural language processing tasks (Devlin et al., 2019; Liu et al., 2019a; Raffel et al., 2020; Brown et al., 2020; Chowdhery et al., 2023) due to their superior capacity in modeling rich in-context semantics and long-range dependencies. Despite this, empirical studies identify two critical flaws in their self-attention mechanisms: attention dilution, where attention weights are diffusely distributed (Fu et al., 2026), and the attention sink effect, where a small set of structurally salient tokens dominate the allocation (Gu et al., 2024; Yu et al., 2024). These flaws reduce the model's sensitivity to task-critical tokens. For instance, in sentiment analysis, vanilla self-attention often assigns disproportionately high weights to neutral function words (e.g., surprisingly", despite") rather than sentiment-bearing ones (e.g., engaging", slow"), impairing fine-grained downstream interpretation.

To mitigate this misallocation, recent attention supervision approaches (Wu et al., 2024; Choi et al., 2020; Xue et al., 2025; Stacey et al., 2022; Zaidan et al., 2007; Choi et al., 2025) regularize self-attention by aligning it with human annotations or static heuristics (Arous et al., 2021; Bao et al., 2018). However, these methods suffer from two fundamental limitations. First, token-level manual annotations are prohibitively costly and heavily domain-confined, restricting cross-task generalizability. Second, since token importance is inherently context-dependent, static labeling lacks the flexibility to provide adaptive guidance. While dynamic, context-aware attention supervision is highly desirable, explicitly tracking token importance poses an inherent combinatorial challenge—even short sequences yield factorial-scale permutations, rendering direct supervision computationally intractable.

To bridge this gap and circumvent the combinatorial explosion, we propose AttCal: a self-supervised, task-space sensitivity-driven attention calibration framework. Our core insight is that token importance can be measured via task-space sensitivity—i.e., how the model's task-aware output responds to localized token perturbations. By formulating attention calibration as a stochastic optimization problem, AttCal operates as an annotation-free, closed-loop system with a fully frozen PLM backbone and a lightweight feedback module. Specifically, token-level attention distributions serve as probabilistic selectors that guide minimal token-deletion perturbations. The feedback module then evaluates the resulting semantic shifts in the task-aware embedding space to generate a sensitivity signal, which iteratively guides the end-to-end align-

ment of attention distributions via policy gradient learning.

Our main contributions are summarized as follows:

(1) Novel Framework: We introduce AttCal, an annotation-free attention calibration framework that aligns attention with task-decisive tokens via task-space sensitivity feedback, eliminating reliance on manual or heuristic labels.

(2) Stochastic Optimization: We reformulate attention allocation as a stochastic policy optimization problem solved via policy gradient learning, effectively resolving the combinatorial complexity of dynamic attention supervision.

(3) Efficient Training: We design an attention-only tuning regime that freezes the PLM backbone, preventing representation drift and significantly cutting computational overhead.

(4) Empirical Validation: Extensive experiments across five standard benchmarks demonstrate that AttCal consistently outperforms state-of-the-art attention supervision baselines in both classification accuracy and human-aligned interpretability.

2 Related work

Recent studies have sought to improve the reliability and interpretability of neural models by supervising attention distributions using human-annotated signals (Stacey et al., 2022; Zhang et al., 2019). For example, (Choi et al., 2020) proposed SANA, leveraging expert annotations to provide token-level supervision, guiding the model to focus on semantically or theoretically important tokens. Similarly, (Chriqui et al., 2025) introduced Human-Machine Attention Learning (HuMAL), aligning model attention with human self-reported attention patterns, while (McGuire and Tomuro, 2021) developed a Human Attention Network (HAN) for Visual Question Answering, generating human-like attention maps from VQA-HAT to supervise model attention. These approaches have demonstrated improvements in both performance and interpretability. However, they rely on costly human annotations and are often restricted to specific domains, limiting their scalability and applicability across diverse tasks. To address these limitations, researchers have explored alternative methods that provide supervisory signals without manual annotation.

To reduce dependency on manual annotations and improve scalability, heuristic-based weak supervision and attention regularization approaches have been proposed. Heuristic-based methods construct supervisory signals from statistical features, syntactic dependencies, or keyword extraction algorithms (Wu et al., 2024), providing approximate guidance to model attention. Another line of work focuses on mitigating attention bias or collapse in long sequences, which can lead to over-concentration on frequent or structurally salient tokens. For instance, (Yang et al., 2021) introduced auxiliary loss functions to penalize low-entropy attention, while (Cheng et al., 2022) proposed Entropy Guided Loss (EGL) to distinguish foreground from background regions in attention maps. These approaches encourage models to allocate attention more evenly and focus on informative tokens, improving task performance and robustness. Nonetheless, their reliance on static or pre-defined signals limits flexibility, preventing adaptation to varying textual contexts and dynamic input scenarios. This highlights the need for more adaptive, context-sensitive attention supervision mechanisms.

Unlike prior attention supervision methods that rely on human annotations, static weak signals, or heuristic priors, our work explores attention optimization from a policy learning perspective. Instead of prescribing token importance explicitly, attention is adjusted based on task-aware feedback derived from model behavior under controlled input perturbations. This formulation allows attention distributions to be refined without external supervision and makes the approach naturally applicable to PLMs.

3 Methodology

3.1 Attention as a Stochastic Policy

In this section, we formalize self-attention distributions as stochastic policies for token deletion. This formalization serves as the core foundation of AttCal's optimization framework.

Consider a Transformer layer l with input $\boldsymbol{X}^l = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{n \times d}$. Here, n denotes sequence length, and d denotes hidden dimension.

For each attention head $h \in [H]$, we compute queries, keys, and values as:

$$\boldsymbol{Q}_h = \boldsymbol{X}^l \boldsymbol{W}_h^Q, \quad \boldsymbol{K}_h = \boldsymbol{X}^l \boldsymbol{W}_h^K, \quad \boldsymbol{V}_h = \boldsymbol{X}^l \boldsymbol{W}_h^V, \tag{1}$$

where $\boldsymbol{W}_h^Q, \boldsymbol{W}_h^K, \boldsymbol{W}_h^V \in \mathbb{R}^{d \times d_h}$ are trainable projection matrices, and $d_h = d/H$ is the per-head hidden dimension.

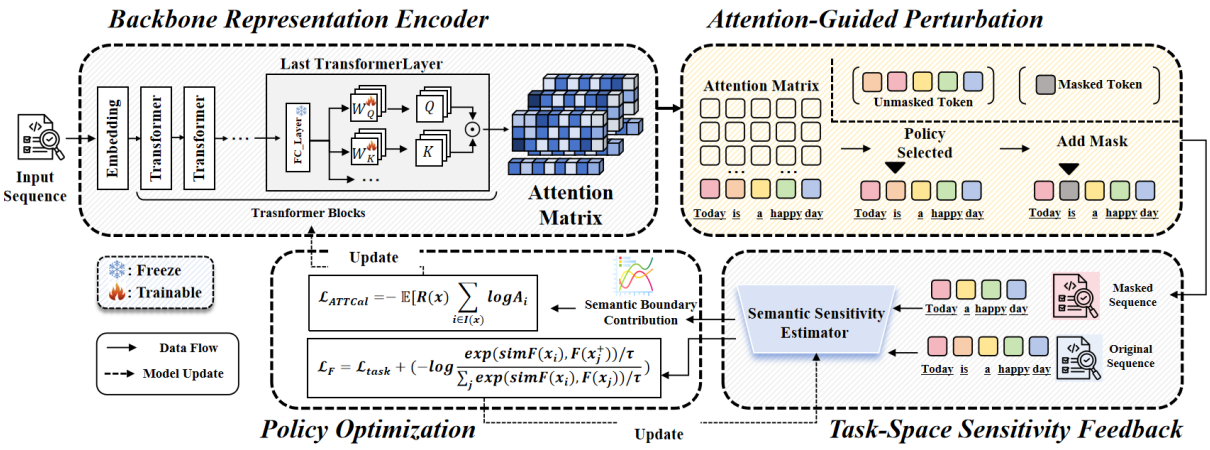

Figure 1: An overview of AttCal. The model first obtains the attention distribution via the Backbone Representation Encoder, followed by masking selected tokens in the Attention-Guided Perturbation module. The Task-Space Sensitivity Feedback module evaluates semantic deviations before and after perturbation to generate reward signals. Finally, the Policy Optimization module updates the attention parameters in the last layer by jointly minimizing $\mathcal{L}_{ATTCal}$ and $\mathcal{L}_F$.

The standard self-attention output is computed as:

$$\text{Attention}(\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h) = \text{softmax}\left(\frac{\boldsymbol{Q}_h \boldsymbol{K}_h^\top}{\sqrt{d_h}}\right)\boldsymbol{V}_h, \tag{2}$$

where $d_h = d/H$.

The resulting attention weights form a valid probability distribution over input tokens. We interpret this normalized attention distribution as a stochastic policy:

$$\pi_\theta(i \mid x) = A_i, \tag{3}$$

where θ denotes trainable attention-only parameters, and A_i is the attention weight of token x_i.

We remove the sampled tokens to construct the perturbed input, formally denoted as $I(x)$. This stochastic token deletion process enables Monte Carlo estimation of semantic boundary sensitivity, with no need for explicit token-level supervision.

3.2 Reward Definition and Credit Assignment

3.2.1 Semantic Boundary Contribution

This section formalizes the semantic boundary contribution, which quantifies the task-level importance of each token. It serves as the core basis for our attention calibration objective.

Let $F(\cdot)$ denote a task-aware semantic projection function implemented by the feedback module(detailed in Section 3.4).

For an input sequence $x = (x_1, ..., x_n)$, we define the semantic boundary contribution of token x_i as

$$B_i(x) = \mathbb{E}_{I_i}\left[d\big(F(x), F(I_i(x))\big)\right], \tag{4}$$

where $I_i(x)$ denotes the stochastic perturbation operator that removes only token x_i from the input sequence, consistent with the perturbation paradigm defined in Section 3.1, and $d(\cdot,\cdot)$ denotes the L2 Euclidean distance metric in the task-aware semantic embedding space;

The expectation $\mathbb{E}_{I_i}$ is estimated via Monte Carlo sampling over multiple perturbation runs. We normalize this contribution to get the latent target distribution for attention calibration:

$$\tilde{B}_i(x) = \frac{\mathbb{E}_{I_i}\left[d\left(F(x), F\left(I_i(x)\right)\right)\right]}{\sum_{j=1}^{n} \mathbb{E}_{I_j}\left[d\left(F(x), F\left(I_j(x)\right)\right)\right]}, \tag{5}$$

This normalized distribution is the implicit target our calibrated attention aims to approximate. However, $\tilde{B}_i(x)$ is not directly observable, and must be estimated via stochastic token deletion perturbations. In practice, we approximate it by measuring the semantic distance $d(F(x), F(I(x)))$ between the original input and its multi-token perturbed counterpart, under the assumption of approximately additive semantic contributions under small perturbation budgets.

We then formulate attention optimization as a policy gradient problem, where attention weights parameterize the stochastic token selection policy defined in Section 3.1, and semantic boundary shifts serve as the reward signal for policy updates.

3.2.2 Semantic Reward from Task-Space Sensitivity

Let $\pi_\theta(i|x) = A_i$ denote the attention-based stochastic policy for token deletion, consistent with the definition in Section 3.2. Given a sampled perturbation $I(x)$ generated by deleting a token subset according to π_θ, we define a scalar semantic sensitivity reward as:

$$R_{sem}(x) = d(F(x), F(I(x))) \tag{6}$$

where $F(\cdot)$ is the task-aware semantic projection function defined in Section 3.2.1, and $d(\cdot, \cdot)$ is the L2 Euclidean distance in the task-aware embedding space.

This reward quantifies the aggregate semantic boundary shift induced by the deleted token subset. Intuitively, tokens critical to the target task will induce a larger displacement in the task-space representation when removed. Thus, deleting more task-critical tokens will yield a higher semantic reward.

Importantly, we define this reward at the perturbation level, not the individual token level. This design avoids heuristic or manually designed token-wise reward decomposition. It also eliminates the need for additional assumptions about token independence or linearity of semantic contributions.

This reward directly enables the estimation of the latent target distribution $\tilde{B}_i(x)$ defined in Section 3.3.1. Since $\tilde{B}_i(x)$ is not directly observable, we approximate it via the semantic distance measured in this reward function. Based on this, we formulate attention calibration as a policy gradient optimization problem, where attention weights parameterize the stochastic token selection policy, and semantic boundary shifts serve as the core reward signal.

3.2.3 Multi-Token Deletion as Monte Carlo Estimation

In AttCal, multiple tokens may be deleted simultaneously in a single perturbation. This multi-token deletion strategy acts as a Monte Carlo estimator for aggregate semantic boundary contribution. It relies on a core assumption: semantic shifts from individual tokens are approximately additive, with weak interactions under small perturbation budgets. Formally, the expected semantic reward under our attention-based stochastic policy can be written as:

$$\mathbb{E}_{I(x)\sim\pi\theta}\left[R_{\text{sem}}(x)\right] \approx \sum_{i=1}^{n} A_i \cdot B_i(x), \tag{7}$$

where π_θ and A_i denote the attention-based stochastic token deletion policy defined in Section 3.1, and$B_i(x)$ is the latent semantic boundary contribution of token x_i, consistent with the definition in Eq. (6), and $Rsem(x)$ is the semantic sensitivity reward defined in Section 3.2.2.

This approximation leverages the additive property of semantic contributions under small perturbations. It enables efficient estimation of individual token importance via batch-level multi-token perturbations.

Under this formulation, we design policy gradient updates as defined in (Sutton and Barto, 2018). These updates encourage the model to assign higher attention weights to tokens with larger semantic boundary contributions. While weak token interactions may exist, the stochastic sampling mechanism ensures consistent reinforcement. Tokens that induce large semantic shifts are consistently reinforced across training iterations. Empirically, we observe stable convergence of attention distributions under this approximation.

3.2.4 Task-Aware Penalty for Stability

Purely maximizing semantic sensitivity may encourage perturbations that excessively degrade task performance. To mitigate this effect, we introduce a task-aware penalty based on the change in task loss. Let $\ell(x)$ and $\ell(I(x))$ denote the task loss before and after perturbation, respectively. The penalty term is defined as

$$R_{\text{task}}(x) = \max\big(0,\ \ell(I(x)) - \ell(x)\big). \tag{8}$$

This hinge-form penalty only takes effect when perturbation increases the task loss, avoiding unnecessary regularization for benign perturbations. We then combine the semantic sensitivity reward and task-aware penalty to form the final reward for attention optimization:

$$R(x) = R_{\text{sem}}(x) - \lambda R_{\text{task}}(x), \tag{9}$$

where λ>0 is a hyperparameter that controls the trade-off between semantic sensitivity and task stability.

This final reward formulation has two core design goals. First, it encourages the model to assign higher attention to tokens that are semantically decisive for the classification task. Second, it discourages perturbations that lead to disproportionate degradation in task performance.

3.2.5 Final Optimization Objective

Based on the final reward defined above, we formulate attention calibration as a policy gradient optimization problem, following the standard REINFORCE framework (Sutton and Barto, 2018).

Recall from Section 3. 1 that we treat the attention distribution $A_i = \pi_\theta(i|x)$ as a stochastic policy for token deletion. The overall optimization objective for attention parameters θ is derived as:

$$\nabla_\theta \mathcal{L}_{\text{PG}} = -\mathbb{E}_{I(x)\sim\pi_\theta}\left[R(x) \cdot \sum_{i\in I(x)} \nabla_\theta \log \pi_\theta(i|x)\right] \tag{10}$$

During the entire optimization process, we keep the backbone PLM fully frozen. Only the self-attention projection parameters θ are updated. This design isolates the pure effect of attention calibration, avoids representation drift, and preserves the pre-trained model's generalization ability. This objective enables AttCal to align attention distributions with latent task-specific semantic boundary contributions. It relies exclusively on model-internal feedback, with no need for external manual annotations or heuristic token importance rules.

3.3 Task-Aware Feedback Module

This module generates task-aware semantic embeddings to compute the semantic sensitivity reward defined in Section 3.2.2. Its core goal is to capture task-relevant semantic changes, while maintaining stability under small input perturbations.

Native frozen PLM representations are optimized for general semantic modeling. They lack sensitivity to task-specific semantic shifts induced by token deletion. We thus design this lightweight task-aware projection module to address this gap.

3.3.1 Semantic Projection Pipeline

Given an input sequence x, the frozen backbone PLM outputs token-level hidden representations $H(x) \in \mathbb{R}^{n\times d}$. Here, n is sequence length, and d is the hidden dimension of the PLM.

We first compute a pooled sequence representation via mean pooling:

$$h(x) = \begin{cases} \frac{\sum_{i=1}^n m_i h_i}{\sum_{i=1}^n m_i}, & \text{if mask } m \text{ is available,} \\ \frac{1}{n}\sum_{i=1}^n h_i, & \text{otherwise,} \end{cases} \tag{11}$$

where h_i denotes the i-th token representation from $H(x)$, and $m_i \in \{0, 1\}$ is the i-th element of the padding mask.

The pooled representation is fed into a lightweight projection network $g(\cdot)$. This network consists of two fully connected layers, with ReLU activation and dropout regularization:

$$z(x) = g(h(x)). \tag{12}$$

Finally, we apply L2 normalization to the output to get the task-aware semantic embedding:

$$F(x) = \frac{z(x)}{\|z(x)\|_2}. \tag{13}$$

Since $F(x)$ is L2-normalized, its L2 distance is monotonically related to cosine similarity. This ensures consistency between our semantic shift measurement and the module's training objectives.

3.3.2 Module Training Objectives

We train the feedback module on the labeled training set of the target text classification task. This avoids data leakage, as we use the same data split for subsequent attention calibration.

We first optimize the module with a standard cross-entropy classification loss. We attach a lightweight linear classification head to the projected representation $z(x)$. The classification loss is defined as:

$$\mathcal{L}_{\text{task}} = -\sum_{(x,y)} y^\top \log \hat{y}, \tag{14}$$

where y is the ground-truth one-hot class label, and $\hat{y}$ is the predicted label distribution derived from $z(x)$. This objective aligns the embedding space with the task's decision boundary.

To further improve the embedding's discriminability, we add an optional contrastive regularization term. We adopt the InfoNCE loss from (Oord et al., 2018). For a batch of samples, instances with the same label are positive pairs, and others are negative pairs. The contrastive loss is:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp\left(\text{sim}(F(x_i), F(x_i^+))/\tau\right)}{\sum_j \exp\left(\text{sim}(F(x_i), F(x_j))/\tau\right)}, \tag{15}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity; x_i^+ is a positive sample sharing the same label as x_i; x_j are other samples in the same batch; τ is a temperature hyperparameter. This objective pulls semantically similar samples closer, and pushes dissimilar samples apart in the embedding space.

The final training objective for the feedback module is:

$$\mathcal{L}_F = \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{con}}, \tag{16}$$

where $\alpha \geq 0$ controls the strength of contrastive regularization. We set $\alpha = 0$ by default, with the contrastive term as an optional enhancement.

3.4 Implementation and Training Details

Attention parameters are optimized using Adam with learning rate 1×10^{-4}. Gradients do not propagate through discrete deletion operations; instead, unbiased gradients are obtained via the score-function estimator.

All models are trained for the same number of steps as their baselines to ensure fair comparison. Freezing the majority of parameters substantially reduces memory consumption and computational cost, making ATTCal scalable to large PLMs.

4 Experiment

4.1 Experimental Setup

To evaluate the effectiveness AttCal framework, we conduct extensive experiments on five representative benchmark datasets:AG News(Zhang et al., 2015), IMDB(Maas et al., 2011), SST-2(Socher et al., 2013), Spam(Almeida and Hidalgo, 2011),HttpParams(Morzeux, 2025).

To assess the general applicability of AttCal on large-scale pre-trained language models, we select several widely-used LLMs as backbone(Team, 2025; AI@Meta, 2024; GLM, 2024; Zhang et al., 2022). For each LLM, we compare its original version with the AttCal-enhanced counterpart, evaluating both Accuracy and Macro F1 scores.

All experiments are conducted on a pre-trained Transformer-based language model. To isolate the effect of attention calibration, the backbone parameters are frozen unless otherwise specified, and only self-attention projection layers are updated according to the strategy described in Section 3. 6.

We compare AttCal against several state-of-the-art attention supervision and de-biasing methods:AS with TF-IDF, SANA:(Choi et al., 2020), DAS(Wu et al., 2024).

In our main experiments, we evaluate the method on various large-scale language models. For comparisons with other attention optimization methods, we adopt RoBERTa-base (Liu et al., 2019b) as the backbone, and all baselines are implemented using the same backbone and training data to ensure a fair comparison.

4.2 Main Results

As demonstrated in Table 1, AttCal consistently achieves significant performance gains across all datasets and base models:Performance Boost: On the AG News dataset, Llama2-7B enhanced with AttCal shows a remarkable accuracy increase of 5. 3% (from 57. 5% to 62. 8%). Sentiment Analysis: For SST-2, Llama3-8B achieves an 15. 8% improvement in accuracy, reaching 84. 0% compared to the baseline. Statistical Significance: Most improvements are statistically significant with $p < 0.01$ under the Wilcoxon signed-rank test, validating the robustness of our method. The improvements are more pronounced on datasets with longer inputs (e. g. , IMDB), suggesting that calibrating attention helps mitigate attention dilution in long-context settings.

Table 2: Results of classification on AGnews, IMDB, SST-2.

Model	AGnews		IMDB		SST2	
	Acc. (%)	Ma-F	Acc. (%)	Ma-F	Acc. (%)	Ma-F
ATT	91.87	0.92	86.28	0.86	90.38	0.90
ATT+AS	89.32	0.89	88.06	0.88	89.32	0.89
SANA	85.12	0.85	82.64	0.83	85.27	0.85
DAS	91.57	0.92	88.32	0.88	90.18	0.90
AttCal	**92.12**	**0.92**	**88.68**	**0.89**	**91.33**	**0.91**

4.3 Analysis of Attention Calibration

4.3.1 Comparative Analysis

Table 2 presents the comprehensive classification performance of our proposed method alongside four strong baselines across three benchmark datasets (AGnews, IMDB, and SST-2). Overall, our proposed AttCal consistently outperforms all baseline models in terms of both Accuracy and Macro-F1 score, demonstrating its superior capability in text classification tasks.

Specifically, on the AGnews dataset, AttCal achieves the highest accuracy of 92.12%, outperforming the strongest baseline, ATT, by a margin of 0.25%. A more pronounced performance leap is observed on the IMDB and SST-2 datasets. On IMDB, AttCal elevates the Accuracy to 88.68% and the Macro-F1 to 0.89, yielding a notable improvement over DAS (88.32% Acc.) and SANA (82.64% Acc.). Similarly, on the SST-2 dataset, our method establishes a new state-of-the-art result with an Accuracy of 91.33% and a Macro-F1 score of 0.91, outclassing the baseline ATT by 0.95% in accuracy.

The systematic superiority of AttCal can be

Table 1: Presents the overall performance comparison on all datasets. Significance: * $p < 0.05$, ** $p < 0.01$ (Wilcoxon Test vs Baseline).

Model	AGnews		IMDB		SST2		Spam		HttpParams	
	Acc. (%)	Ma-F	Acc. (%)	Ma-F	Acc. (%)	Ma-F	Acc. (%)	Ma-F	Acc. (%)	Ma-F
Qwen2.5-7B	80.9±0.4	0.81±0.02	91.3±0.4	0.90±0.01	71.7±0.6	0.71±0.03	97.8±0.2	0.95±0.01	97.5±0.2	0.98±0.01
Qwen2.5-7B(AttCal)	90.1±0.3**	0.90±0.01**	93.1±0.3**	0.92±0.01*	85.3±0.4**	0.84±0.02**	99.6±0.1**	0.99±0.00**	99.7±0.1**	0.99±0.00**
Qwen3-8B	83.1±0.4	0.85±0.02	92.6±0.3	0.93±0.01	86.3±0.3	0.85±0.02	98.5±0.2	0.98±0.00	97.6±0.2	0.97±0.01
Qwen3-8B(AttCal)	86.6±0.3**	0.86±0.01*	94.3±0.2**	0.94±0.01*	87.2±0.3*	0.87±0.01*	98.6±0.1*	0.99±0.00*	99.8±0.1**	0.99±0.00**
Llama2-7B	57.5±0.6	0.56±0.03	72.7±0.5	0.72±0.03	68.2±0.6	0.68±0.03	95.6±0.3	0.91±0.02	97.7±0.2	0.98±0.01
Llama2-7B(AttCal)	62.8±0.5**	0.61±0.02**	73.8±0.4**	0.74±0.02**	83.2±0.5**	0.83±0.02**	99.3±0.1**	0.98±0.01**	99.4±0.1**	0.99±0.00**
Llama3-8B	74.1±0.4	0.77±0.02	84.3±0.4	0.84±0.02	75.2±0.6	0.74±0.03	92.7±0.3	0.94±0.02	96.9±0.3	0.97±0.01
Llama3-8B(AttCal)	86.7±0.3**	0.87±0.01**	89.5±0.3**	0.89±0.01**	84.0±0.5**	0.84±0.02**	99.1±0.2**	0.98±0.01**	99.6±0.1**	0.99±0.00**
ChatGLM3-6B	81.7±0.4	0.82±0.02	79.7±0.5	0.78±0.03	72.3±0.6	0.72±0.03	97.0±0.3	0.97±0.01	97.1±0.3	0.96±0.01
ChatGLM3-6B(AttCal)	90.8±0.3**	0.90±0.01**	83.7±0.4**	0.81±0.02*	85.2±0.5**	0.82±0.02**	98.5±0.2**	0.98±0.01*	99.4±0.1**	0.99±0.00**
ChatGLM4-9B	87.2±0.3	0.87±0.01	89.2±0.3	0.90±0.01	86.8±0.3	0.87±0.01	98.6±0.1	0.98±0.00	97.7±0.2	0.97±0.01
ChatGLM4-9B(AttCal)	87.7±0.3*	0.88±0.01*	93.5±0.3**	0.93±0.01**	87.8±0.3*	0.88±0.01*	98.8±0.1*	0.99±0.00*	99.3±0.1**	0.99±0.00**
OPT-6.7B	84.6±0.4	0.84±0.02	90.6±0.3	0.91±0.01	84.0±0.3	0.84±0.02	97.2±0.2	0.97±0.01	96.6±0.3	0.94±0.02
OPT-6.7B(AttCal)	87.5±0.3**	0.87±0.01**	93.3±0.3**	0.93±0.01**	86.8±0.3**	0.87±0.01**	98.2±0.2**	0.98±0.01*	98.8±0.2**	0.99±0.01**

attributed to its advanced attention calibration mechanism. While baseline methods like ATT and ATT+AS occasionally suffer from attention misalignment—where the model inadvertently focuses on uninformative tokens or background noise—AttCal effectively calibrates the attention distribution. This ensures that the model dynamically prioritizes highly concentrated, semantically rich tokens that are crucial for classification. The consistent gains across diverse text styles, ranging from long-form movie reviews (IMDB) to short sentences (SST-2), strongly validate the robustness and generalizability of our approach.

Table 3: Performance evaluation on the AG News dataset using Qwen3-8B with various window sizes (k).

k	ACC (%)	PRE (%)	REC (%)	F1
3	86.20	86.38	86.04	86.21
5	85.30	86.16	85.44	85.80
10	83.00	84.40	83.18	83.79
15	81.30	84.03	81.70	82.85
20	82.30	81.89	82.43	81.66

4.3.2 Hyperparameter Sensitivity Analysis

To investigate the impact of the window size k on model performance, we conduct a sensitivity analysis on the AG News dataset using Qwen3-8B, varying k from 3 to 20. As illustrated in Table 3, the model exhibits a unique non-monotonic performance trend. Specifically, the optimal performance is achieved at a tight window size of $k = 3$, yielding an Accuracy of 86.20% and an F1 score of 86.21%. As k expands from 3 to 15, we observe a monotonic degradation in all evaluation metrics, with the F1 score dropping to its lowest point of 82.85% at $k = 15$. This performance decay suggests that an excessively large window size initially introduces redundant linguistic noise and dilutes the model's focus on highly concentrated, locally informative attention heads.Interestingly, when k is further extended to 20, the performance exhibits a noticeable rebound, with the Accuracy and F1 score recovering to 82.30% and 81.66%, respectively. This inflection point implies that while intermediate window sizes suffer from a poor signal-to-noise ratio, a sufficiently large context window ($k = 20$) allows the model to capture broader, long-range semantic dependencies that compensate for the local noise. Overall, the results demonstrate that a compact window ($k = 3$) strikes the best balance between filtering out irrelevant context and capturing crucial local semantics for news classification.

4.4 Ablation Study

4.4.1 Effect of Attention Gradient Updates

This section investigates whether the observed improvements stem from attention calibration rather than perturbation-based training or additional optimization effects. We consider four model variants to disentangle the effect of attention calibration from perturbation-based training. Baseline denotes standard fine-tuning without perturbation or attention updates. Perturbation Only applies attention-guided token masking while freezing attention parameters. Random Perturbation performs token masking uniformly at random. AttCal (Full) jointly performs attention-guided perturbation and gradient-based optimization of attention parameters

Table 4: Ablation results of AttCal on RoBERTa-base.

Model Variant	AG News (%)	IMDB F1	SST-2 (%)
Baseline	83.1	0.83	86.3
Perturbation Only	83.3	0.83	86.5
Random Perturbation	82.9	0.82	85.9
AttCal (Full)	**84.8**	**0.85**	**87.0**

Table 4 reports results on AG News, IMDB, and SST-2 under four controlled variants. Compared to the baseline RoBERTa-base model, applying attention-guided perturbations without updating attention parameters ("Perturbation Only") yields only marginal improvements (e.g. , +0. 2% on AG News and +0.2% on SST-2), while random perturbations consistently degrade performance across all datasets. In contrast, the full AttCal configuration, which optimizes attention parameters via task-aware feedback, achieves substantial and consistent gains, improving AG News accuracy by +1.7%, IMDB F1 by +0.02, and SST-2 accuracy by +0.7% over the baseline.

4.4.2 Effect of Head Aggregation Strategies

Table 5: Ablation study on head aggregation strategies across five benchmarks. Avg. Pooling and Entropy-wt. denote Average Pooling and Entropy-weighted strategies, respectively. Accuracy (%) is reported, and the best results are highlighted in bold.

Dataset	**Avg. Pooling**	**Entropy-wt. (Ours)**	**Improvement**
AG News	80.1	**84.8**	*+4.7*
IMDB	81.5	**85.0**	*+3.5*
SST-2	82.6	**87.0**	*+4.4*
HttpParam	96.2	**98.5**	*+2.3*
Spam	95.8	**98.2**	*+2.4*

To evaluate the efficacy of our proposed entropy-weighted aggregation mechanism, we conduct an ablation study by comparing it against the standard average pooling baseline across all five benchmarks. As summarized in Table **??**, our entropy-weighted strategy consistently outclasses average pooling by a convincing margin, achieving substantial accuracy improvements ranging from +2.3% to +4.7%. Notably, the most pronounced gains are observed on the AG News and SST-2 datasets, where the classification accuracy surges by 4.7% (84.8% vs. 80.1%) and 4.4% (87.0% vs. 82.6%), respectively.

This comprehensive superiority stems from the intrinsic limitation of average pooling, which treats all attention heads uniformly and inadvertently distorts the final representation by pooling diluted, noisy attention maps alongside informative ones. Conversely, by leveraging entropy as a parameter-free proxy, our method successfully quantifies the concentration of information within each head. It strategically dynamically upweights highly focused, informative attention heads while penalizing dispersed and uninformative ones. The consistent performance leap across both text classification and security-related benchmarks (e.g., HttpParam and Spam) strongly vindicates that filtering attention noise via entropy weighting is critical for robust feature fusion.

5 Conclusion

In this work, we present AttCal, an annotation-free attention calibration framework that mitigates attention dilution and sink effects by leveraging dynamic, task-space sensitivity feedback. By optimizing attention distributions via policy gradient learning while freezing the PLM backbone, AttCal adaptively captures context-dependent token importance without relying on costly human annotations or rigid heuristics. Extensive experiments on AG-News, IMDB, and SST-2 demonstrate that AttCal consistently outperforms vanilla PLMs and state-of-the-art attention supervision baselines, achieving substantial classification accuracy gains and delivering highly interpretable, human-aligned attention maps. Rationality analyses and hyperparameter studies further confirm its stability and robustness. Future work will explore extending AttCal to more complex scenarios, including multimodal learning and real-world log analysis, to further enhance model reliability and robustness in high-stakes domains.

6 Limitations

While effective, our method has several limitations:

1) The proposed perturbation-based sensitivity estimation may be less discriminative for highly redundant texts, where removing a small number of tokens does not substantially alter task predictions due to semantic overlap.

2) The framework depends on the perturbation budget k. Very small perturbations may yield weak feedback signals, while overly large perturbations can introduce excessive semantic distortion. Although we observe robustness within a reasonable range of k, its selection remains task-dependent.

3) Our approach assumes a stable task prediction head, as task-space sensitivity is derived from

changes in output distributions. Instability or poor calibration of the task head may introduce noise into the feedback signal.

These limitations reflect the inherent trade-offs of perturbation-based attention calibration and suggest directions for future refinement.

References

AI@Meta. 2024. Llama 3 model card.

Tiago Almeida and Jos Hidalgo. 2011. SMS Spam Collection. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5CC84.

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. Marta: Leveraging human rationales for explainable text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 5868–5876.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yi Cheng, Ying Sun, Hehe Fan, Tao Zhuo, Joo-Hwee Lim, and Mohan Kankanhalli. 2022. Entropy guided attention network for weakly-supervised action localization. *Pattern Recognition*, 129:108718.

Daewon Choi, Jimin Lee, Jihoon Tack, Woomin Song, Saket Dingliwal, Sai Muralidhar Jayanthi, Bhavana Ganesh, Jinwoo Shin, Aram Galstyan, and Sravan Babu Bodapati. 2025. Think clearly: Improving reasoning via redundant token pruning. *arXiv preprint arXiv:2507.08806*.

Seungtaek Choi, Haeju Park, Jinyoung Yeo, and Seungwon Hwang. 2020. Less is more: Attention supervision with counterfactuals for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6695–6704.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Avihay Chriqui, Inbal Yahav, Dov Teeni, and Ahmed Abbasi. 2025. Aligning human and machine attention for enhanced supervised learning. *arXiv preprint arXiv:2502.06811*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Zichuan Fu, Wentao Song, Guojing Li, Yejing Wang, Xian Wu, Yimin Deng, Hanyu Yan, Yefeng Zheng, and Xiangyu Zhao. 2026. Attention needs to focus: A unified perspective on attention allocation. *arXiv preprint arXiv:2601.00919*.

Team GLM. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2024. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Erik McGuire and Noriko Tomuro. 2021. Sentiment analysis with cognitive attention supervision. In *Canadian AI*.

Morzeux. 2025. Httpparamsdataset. Accessed: 2025-09-14.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank.

In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11349–11357.

Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. MIT Press.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yiquan Wu, Yifei Liu, Ziyu Zhao, Weiming Lu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2024. De-biased attention supervision for text classification with causality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19279–19287.

Fanfu Xue, Jiande Sun, Yaqi Xue, Qiang Wu, Lei Zhu, Xiaojun Chang, and Sen-Ching Cheung. 2025. Attention guidance by cross-domain supervision signals for scene text recognition. *IEEE Transactions on Image Processing*.

Sen Yang, Zhicheng Wang, Ze Chen, Yanjie Li, Shoukui Zhang, Zhibin Quan, Shu-Tao Xia, Yiping Bao, Erjin Zhou, and Wankou Yang. 2021. Attend to who you are: Supervising self-attention for keypoint detection and instance-aware association. *arXiv preprint arXiv:2111.12892*.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. 2024. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. 2019. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 ieee winter conference on applications of computer vision (wacv)*, pages 349–357. IEEE.

A Appendix

A.1 Attention-Only Calibration vs. Full Fine-Tuning

We compare attention-only calibration with conventional full fine-tuning to evaluate the efficiency–performance trade-off of AttCal. While full fine-tuning updates all backbone parameters, AttCal restricts optimization to the attention projection layers (Q/K/V and output), leaving the remaining parameters frozen.

As shown in Table 6, AttCal updates only 3. 2% of the parameters of RoBERTa-base, resulting in a 3× reduction in per-epoch training time and more than 2× lower GPU memory consumption. Despite this substantial reduction in computational cost, AttCal achieves competitive performance compared to full fine-tuning. These results demonstrate that effective task adaptation does not require full parameter updates. By calibrating attention distributions alone, AttCal offers a lightweight and scalable alternative to full fine-tuning, particularly suitable for resource-constrained or large-scale deployment scenarios.

A.2 Human Alignment Verification

A case study on the AG News dataset highlights the effectiveness of AttCal in addressing the limitations of standard attention mechanisms. As shown in Figure 2, the baseline attention without supervision tends to concentrate on function or auxiliary words (e. g. , in, has, ever), which dilutes the information flow and misguides the model into predicting the wrong label (World). This illustrates the problems of information dilution and incorrect allocation inherent in conventional Transformers. In contrast, AttCal provides adversarial feedback that encourages the model to reallocate attention toward semantically critical tokens such as Gunners, Manchester, and Liverpool. By doing so, AttCal not only improves classification accuracy but also enhances interpretability, as the attention distribution aligns more closely with human intuition about relevant cues. This case demonstrates that AttCal can mitigate the weaknesses of global self-attention and offer more reliable and transparent decision-making in text classification tasks.

To ensure reliable and interpretable human annotations of critical tokens for sentiment classification, we design a structured annotation protocol. Three annotators with backgrounds in computational linguistics or NLP are recruited and trained

Figure 2: Case study on the AG News dataset. The color intensity indicates attention values: darker colors correspond to higher attention weights.

on the task. They should identify the top-3 tokens in each sentence that are most semantically influential in determining sentiment polarity—focusing on content words (nouns, verbs, adjectives, adverbs) that carry affective meaning or modulate sentiment intensity. Function words (e.g., “the”, “and”) are only considered critical if they significantly alter sentiment interpretation (e. g., negation: “not”, intensifiers: “very”, “extremely”). Annotators are instructed to avoid selecting tokens based on frequency or position alone, and instead justify their choices using semantic reasoning.

For the 100 randomly sampled sentences from the SST-2 development set, each annotator independently labels the top-3 critical tokens per sample. Disagreements are resolved via consensus discussion, where annotators review conflicting selections and reach agreement based on shared linguistic intuition and in-context analysis. Inter-annotator agreement is measured using Fleiss’ kappa (kappa = 0.64), indicating substantial agreement, which supports the reliability of the final annotated set. The resulting human-labeled token sets serve as ground truth for evaluating attention alignment.

We evaluate the interpretability of attention distributions by measuring their alignment with human annotations using two metrics: Intersection-over-Union (IoU) to assess overlap between the model’s top-3 attended tokens and human-identified critical tokens, and Spearman’s rank correlation (rho) to measure agreement in importance ordering. As shown in Table 7, our proposed AttCal method achieves significantly higher alignment than baseline approaches: an IoU of 0. 68 and rho = 0. 72, both statistically significant ($p < 0.01$). This improvement suggests that AttCal not only enhances performance but also produces

Table 6: Comparison of training strategies on RoBERTa-base. Full Fine-tuning: all parameters; AttCal (Attention-only): only attention projection layers (Q/K/V/Output). Param. ratio is relative to RoBERTa-base. Time and GPU memory measured on a single A100 GPU on the AG News dataset. Accuracy and F1 results are discussed in the main text.

Training Strategy	Param. Ratio (%)	Time (h/epoch)	GPU Mem (GB)
Full Fine-tuning	100.0	0.3	10
AttCal (Attention-only)	3.2	0.1	4

Table 7: Alignment between model attention and human-annotated critical tokens. IOU is the intersection-over-union of top-3 model attention tokens and human-annotated tokens. Spearman's ρ measures the rank correlation of importance order. Significance: ** p<0.01.

Method	IOU (Human Alignment)	Spearman's ρ
ATT	0.43	0.51
SANA	0.52	0.58
DAS	0.57	0.63
AttCal	0.68**	0.72**

more human-aligned attention patterns, indicating increased interpretability and rationality in focus allocation. In contrast, baselines such as ATT, SANA, and DAS exhibit weaker alignment, particularly in ranking consistency (e. g. , rho = 0. 51 for ATT), highlighting their tendency to assign attention to salient but semantically less critical tokens. These results demonstrate that our method effectively recalibrates attention toward task-relevant semantics, thereby improving model interpretability.

A.3 Necessity Verification

Table 8: Semantic similarity reduction after masking the top-3 predicted important tokens. Sim↓ denotes the reduction rate of semantic similarity.

Dataset	ATT	SANA	DAS	AttCal
AGNews	12. 5	18. 0	20. 5	27. 0
IMDB	14. 0	20. 5	19. 8	25. 5
SST-2	13. 5	19. 0	21. 0	24. 8
Average (Sim↓, %)	13. 3	19. 2	20. 4	25. 8

We evaluate the necessity of the identified tokens by masking the top-3 most important tokens (as ranked by each method) and measuring the resulting reduction in semantic similarity between the original and masked sentence embeddings, denoted as Sim_{-}. As shown in Table 8, our AttCal method achieves a 27. 0% reduction on AGNews, significantly higher than the 12. 5% from vanilla attention, indicating that AttCal identifies tokens whose removal causes greater disruption to the sentence's semantic content. This suggests that the learned attention under AttCal is more sensitive to semantically salient components.

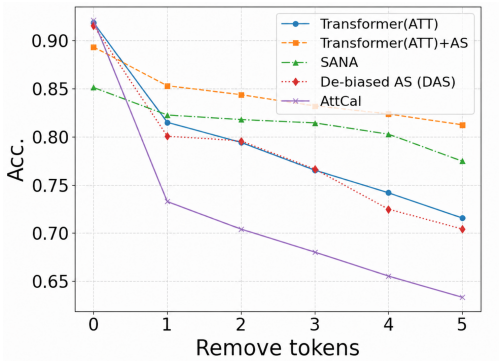

Figure 3: Performance degradation under token deletion on the AGNews dataset. Tokens were ranked by attention weights, and the top-N tokens were sequentially removed.

In addition, we conduct token deletion experiments on the AGNews dataset, progressively removing the top-N tokens ranked by attention weights, and measure performance degradation (Figure 3). The model trained with AttCal exhibits a sharp initial drop in accuracy when the top token is removed, followed by a continued decline—consistent with the behavior of a model whose attention is focused on a small set of highly influential tokens. While this steeper early drop may reflect higher sensitivity to key terms, it also highlights that AttCal's attention distribution is less uniformly spread across the input, potentially leading to more interpretable and task-relevant focus compared to baselines like SANA or DAS.

Notably, while AttCal shows faster degradation, its final accuracy at N=5 remains competitive with DAS and superior to ATT, suggesting that the critical tokens are not only well-identified but also contribute meaningfully to the overall prediction.

Table 9: Prediction accuracy when retaining only the top-5% most influential tokens selected by each method.

Dataset	ATT	SANA	DAS	AttCal
AGNews	54. 3	58. 7	61. 2	64. 8
IMDB	60. 5	64. 9	67. 3	70. 1
SST-2	70. 8	73. 4	75. 6	77. 9
Average (Acc, %)	61. 9	65. 7	68. 0	70. 9

A.4 Sufficiency Verification

To assess the sufficiency of the tokens identified by different attention mechanisms, we conduct a token ablation study: for each input sequence, we retain only the top-5% most influential tokens as ranked by the respective attention method (ATT, SANA, DAS, or our AttCal), and replace all other tokens with a [MASK] token. The model then makes predictions based solely on these sparse inputs. As shown in Table 9, models using attention weights from our AttCal framework consistently achieve the highest accuracy across all three datasets, demonstrating that AttCal effectively identifies tokens that are genuinely informative for the downstream task. Notably, on the short-text sentiment benchmark SST-2, AttCal retains 77. 9% accuracy with only five tokens—close to the performance of the full-input model—indicating its ability to pinpoint decisive sentiment cues (e. g. , strong opinion words). On longer and more complex texts such as IMDB reviews and AGNews articles, where in-context nuance plays a larger role, performance naturally declines, yet AttCal still outperforms all baselines by a clear margin (70. 1% on IMDB and 64. 8% on AGNews). This consistent improvement validates that the task-space sensitivity feedback in AttCal yields attention distributions better aligned with true functional importance than heuristic or gradient-free alternatives.