# SMIC: SHARED MULTISCALE INPUT-REFERENCE CONDITIONING FOR NATURAL LANGUAGE INFERENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Natural Language Inference models can be sensitive to localized lexical changes. We introduce Shared Multiscale Input-Reference Conditioning (SMIC), which constructs a single input-dependent multiscale reference and reuses it across Transformer layers. Each layer applies its own multiscale transformation to the shared reference and injects the resulting conditioning feature before the attention update. We instantiate SMIC in a token–context Transformer (TCT) and conduct from-scratch comparisons against decoder-based and Multi-Layer Perceptron (MLP) baselines across all three ANLI rounds, with encoder comparisons on ANLI-R1. TCT-SMIC attains the highest observed ten-seed mean accuracy among the configurations in the primary baseline comparisons on all three ANLI rounds. The ablations further support the contribution of SMIC. We evaluate a real-valued Llama-style plug-in and extend deletion/insertion diagnostics to pretrained ModernBERT encoder and Llama decoder models with and without SMIC. Across the evaluated perturbation protocols, TCT-SMIC exhibits comparatively low pooled-attention redistribution and smaller signed accuracy changes. We further establish a finite-sample bound on average subset-mass redistribution in terms of the mean Jensen–Shannon divergence between the restricted and renormalized attention distributions.

## 1 INTRODUCTION

Scaling Transformer-based language models increases training and inference costs, while larger parameter counts do not always produce proportional gains under comparable optimization budgets (Kaplan et al., 2020; Yao et al., 2022; Petty et al., 2024). Meanwhile, Natural Language Inference (NLI) remains challenging, particularly on ANLI, an adversarial benchmark requiring models to distinguish entailment, contradiction, and neutrality (Nie et al., 2020). Previous work has also shown that NLI predictions can be sensitive to localized lexical changes (Glockner et al., 2018; Chen et al., 2021). This motivates architectures that process localized lexical information without relying solely on larger models.

A sparse lexical replacement modifies only a small subset of the input tokens, but the resulting representation changes can propagate through successive model layers. These changes are not necessarily harmful: some lexical substitutions should alter the predicted NLI relation, whereas others should not. We therefore use perturbation sensitivity as a protocol-specific diagnostic of model behavior under these replacement conditions.

Certified robustness methods address word substitutions (Jia et al., 2019; Wang et al., 2021a; Zhao et al., 2022; Zeng et al., 2023; Zhang et al., 2024b; Ye et al., 2023) and broader text edits (Huang et al., 2023; Zhang et al., 2024a). Empirical studies examine model sensitivity (Wang et al., 2023a) and defenses based on adversarial training (Dong et al., 2021), detection and repair (Zhou et al., 2021), or ensembling, augmentation, contrastive learning, and masked language modeling (Wang et al., 2021b; Yang et al., 2025; Meng et al., 2022; Wang et al., 2023b); explanation-based prompting has also been studied for adversarial NLI (Kavumba et al., 2023). Convolutional sequence models capture local and hierarchical structure (Kim, 2014; Dauphin et al., 2017; Gehring et al., 2017), while feature conditioning modulates intermediate representations using auxiliary information (Perez et al., 2018). Complex-valued networks model amplitude and phase (Lee et al., 2022); EulerFormer uses polar-form embeddings and complex rotations (Tian et al., 2024), and the Complex–Vector



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

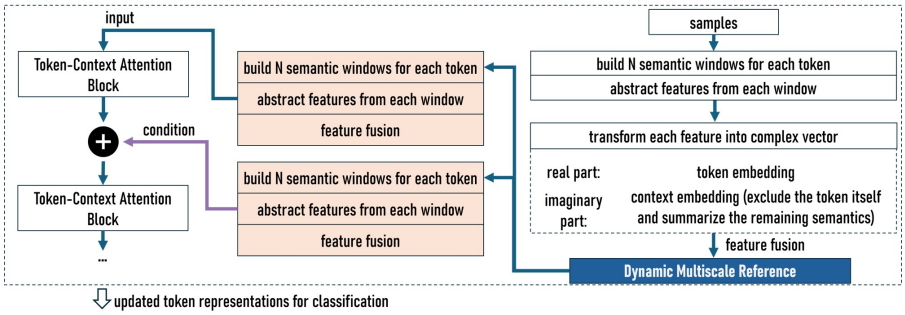



Figure 1: Overview of TCT-SMIC.

Token Representation assigns token and context roles to its Cartesian components (Zhang & Sheng, 2026). However, these studies do not directly examine whether multiscale features extracted from the original input and reused to condition successive Transformer layers can improve NLI accuracy and reduce empirical sensitivity to localized lexical changes.

Inspired by multiscale reference-based processing in signal processing (Martin, 2001; Cohen, 2003; Wang et al., 2017), we introduce Shared Multiscale Input-Reference Conditioning (SMIC), which constructs an input-dependent reference from multiscale convolutional features. Building on the signal view of token embeddings introduced by Zhang & Sheng (2026), TCT-SMIC maps scale-specific features to complex–vector token representations and fuses them into the shared reference. Each layer applies its own multiscale transformation to this reference and injects the resulting features into the token and context components before the Token–Context Attention (TCA) update. Figure 1 summarizes the TCT-SMIC architecture.

**Contributions.** **First**, we introduce SMIC as an architecture-level approach designed to improve ANLI accuracy and reduce protocol-specific sensitivity to sparse lexical replacements without relying solely on increased model scale. **Second**, we provide finite-sample interpretations of signed accuracy change and conditional pooled-attention Jensen–Shannon divergence. We express signed accuracy change as the difference between break and recovery rates and show that mean Jensen–Shannon divergence bounds average subset-mass redistribution within the restricted and renormalized pooled-attention support. **Third**, we compare TCT-SMIC with decoder-based and MLP baselines on all three ANLI rounds and with encoder baselines on ANLI-R1, obtaining the highest observed ten-seed mean accuracy in these comparisons. Ablations assess SMIC's design choices, while fixed-length token and surface-text perturbations show comparatively low attention redistribution and smaller signed accuracy changes. A real-valued Llama-style plug-in and deletion/insertion diagnostics on pretrained ModernBERT and Llama backbones evaluate SMIC across real-valued and pretrained architectures.

## 2 Preliminaries

We build on the complex–vector token representation (CVTR) and Token–Context Attention (TCA) introduced by Zhang & Sheng (2026). For each embedding coordinate, CVTR computes a sequence-level magnitude and assigns token-dependent phases that encode each token's relation to this magnitude. Given an embedding matrix $W = [w_1; \ldots; w_n] \in \mathbb{R}^{n \times d}$, conversion to Cartesian coordinates yields $Z_j = T_j + iC_j \in \mathbb{C}^d$, where $T_j = \Re(Z_j)$ and $C_j = \Im(Z_j)$ are the coupled real-valued token and context components, respectively. Appendix A provides the Cartesian-form derivation.

## 3 Method

TCT-SMIC constructs a complex-valued multiscale reference for each premise–hypothesis pair and reuses it across all attention layers. The reference is input-dependent, whereas its transformations and injection weights are layer-specific.

### 3.1 SHARED MULTISCALE INPUT REFERENCE

Let $u \in \{p, h\}$ denote the premise and hypothesis. For each $u$, let $E^u \in \mathbb{R}^{T_u \times d}$ denote the padded embedding tensor, let $m^u \in \{0,1\}^{T_u}$ denote the active-position mask, and let $n_u = \sum_{j=1}^{T_u} m_j^u$ be the active length. We project the two sequences separately: $X_j^u = m_j^u \left(W_u E_j^u + b_u\right)$. The premise branch uses the candidate kernel set $\mathcal{K}_p = \{1, 3, 7, 13, 21, 33\}$, whereas the hypothesis branch uses $\mathcal{K}_h = \{1, 3, 5, 7, 9, 13\}$. Natural-language dependencies are generally local: related words tend to occur at shorter linear distances (Futrell et al., 2015). We therefore use denser coverage of short receptive fields while retaining larger kernels for long-range interactions. Because premises have a larger maximum length, their kernel bank covers a wider range than the hypothesis bank. Kernel sizes are fixed, while the filters, effective centers, and mixture weights are learned jointly with the model.

For each $k_b \in \mathcal{K}_u$, SMIC computes the length-preserving feature $C_{u,b} = \mathrm{ReLU}\left(\mathrm{Conv}_{u,k_b}\left(X^u\right)\right)$. The trainable $k = 1$ convolution is initialized as an identity mapping and forms the token-local branch $B_{u,0} = C_{u,k=1}$. The default *2LW* configuration adds two learnable Gaussian-mixture branches. For branch $s \in \{1, 2\}$, the receptive-field ratio is $r_{u,s} = \mathrm{clip}\left(a_{u,s} + 0.5 \tanh\left(\theta_{u,s}\right), 10^{-3}, 1 - 10^{-3}\right)$. The premise anchor ratios are $(a_{p,1}, a_{p,2}) = (0.17, 0.33)$, whereas the hypothesis anchor ratios are $(a_{h,1}, a_{h,2}) = (0.30, 0.65)$. These anchors initialize the receptive-field ratios, while the offset parameters are initialized as $\theta_{u,s} \sim \mathcal{N}\left(0, 0.5^2\right)$, and transformed by the bounded $\tanh$ term to determine the trainable effective ratios.

For a minibatch of size $B$, define the mean active length as $n_u^{\mathrm{batch}} = \frac{1}{B} \sum_{q=1}^{B} n_u^{(q)}$. Each module initializes its running statistic with the first minibatch and subsequently updates it as $\overline{n}_u \leftarrow 0.95\overline{n}_u + 0.05 n_u^{\mathrm{batch}}$. The statistic is updated during training and fixed during evaluation. We set the target kernel center at $\mu_{u,s} = \mathrm{clip}\left(r_{u,s}\overline{n}_u, \min \mathcal{K}_u, \max \mathcal{K}_u - 0.5\right)$. We fix the Gaussian bandwidth at $\sigma = 5$. The first mixture excludes the isolated $k = 1$ candidate, $\mathcal{K}_{u,1} = \mathcal{K}_u \setminus \{1\}$, while the second uses the full bank, $\mathcal{K}_{u,2} = \mathcal{K}_u$. For $s \in \{1, 2\}$ and $k_b \in \mathcal{K}_{u,s}$, the normalized weight is $\gamma_{u,s,b} = \frac{\exp\left(-\frac{(k_b - \mu_{u,s})^2}{2\sigma^2}\right)}{\sum_{k_c \in \mathcal{K}_{u,s}} \exp\left(-\frac{(k_c - \mu_{u,s})^2}{2\sigma^2}\right)}$. The resulting branch feature is $B_{u,s} = \sum_{k_b \in \mathcal{K}_{u,s}} \gamma_{u,s,b} C_{u,b}$. The global branch mean-pools the segment tensor and applies a learned linear projection: $g_u = \frac{1}{T_u} \sum_{j=1}^{T_u} X_j^u$, $B_{u,3,j} = W_{u,\mathrm{global}} g_u + b_{u,\mathrm{global}}$. Each branch is then mapped independently to a CVTR as $Q_{u,s} = \Phi_{\mathrm{CVTR}}\left(B_{u,s}\right)$.

Let $c_{u,s}$ be the segment-level fusion logits. The corresponding weights are $\omega_{u,s} = \frac{\exp(c_{u,s})}{\sum_{v=0}^{3} \exp(c_{u,v})}$. We define the shared reference as $R^u = \sum_{s=0}^{3} \omega_{u,s} Q_{u,s} + P^u$, where $P^u$ is a real-valued sinusoidal positional encoding, zeroed at padded positions. Zero-initialized fusion logits assign equal initial weights to the four branches. We concatenate the premise and hypothesis references as $Z^{(0)} = \left[R^p; R^h\right]$. Kernel grids, anchors, and bandwidth remain fixed, while effective ratios, Gaussian-mixture weights, and fusion weights are learned.

### 3.2 LAYER-SPECIFIC CONDITIONING

Each layer applies separate multiscale convolutional modules to the real and imaginary reference components. All modules use the candidate kernel set $\mathcal{K} = \{1, 3, 7, 13, 21, 33\}$, anchor initialization $(0.17, 0.33)$, and fixed bandwidth $\sigma = 5$. Each module contains an isolated $k = 1$ branch, two Gaussian-mixture branches, and one global branch. The candidate kernel sizes are fixed across layers, while the convolutional filters, receptive-field offsets, running length statistics, and branch-fusion logits are learned separately for each layer and component. Because these modules process the concatenated premise–hypothesis reference, both sequences share the same candidate kernel bank. Let $H_r^{(\ell,s)}$ denote branch $s$ produced by the layer-$\ell$ module for component $r \in \{\mathrm{real}, \mathrm{imag}\}$. The layer-specific branch weights are $\beta_r^{(\ell,s)} = \frac{\exp\left(b_r^{(\ell,s)}\right)}{\sum_{v=0}^{3} \exp\left(b_r^{(\ell,v)}\right)}$. The logits $b_r^{(\ell,s)}$ are initialized to zero, giving a uniform initial mixture. The fused conditioning feature is $\widehat{H}_r^{(\ell)} = \sum_{s=0}^{3} \beta_r^{(\ell,s)} H_r^{(\ell,s)}$.

Table 1: ANLI accuracy and training FLOPs across parameter-comparable configurations.

| Model | R1 Acc/FLOPs | Diff [95% CI] | R2 Acc/FLOPs | Diff [95% CI] | R3 Acc/FLOPs | Diff [95% CI] |
|---|---|---|---|---|---|---|
| **0.8B configurations** | | | | | | |
| **TCT-SMIC (0.80B)** | **39.03**/1.06k | – | **44.89**/2.54k | – | **43.17**/4.27k | – |
| Phi-1.5-style (0.82B) | 36.88/1.37k | 2.15 [0.88,3.42] | 41.13/2.60k | 3.76 [2.67,4.85] | 41.59/8.04k | 1.58 [0.83,2.32] |
| Llama-3-style (0.85B) | 36.61/1.22k | 2.42 [0.86,3.98] | 41.12/2.30k | 3.77 [2.52,5.02] | 41.23/5.11k | 1.94 [0.92,2.96] |
| Pythia-style (0.81B) | 36.02/1.23k | 3.01 [1.95,4.07] | 38.17/3.00k | 6.72 [5.16,8.28] | 40.44/3.00k | 2.73 [1.33,4.12] |
| MLP (0.80B) | *37.57*/0.99k | 1.46 [0.25,2.67] | *43.43*/3.42k | 1.46 [0.42,2.50] | *41.83*/6.18k | 1.34 [0.29,2.40] |
| **1B configurations** | | | | | | |
| Phi-1.5-style (1.3B) | *36.76*/2.21k | 2.27 [0.94,3.60] | 40.75/4.34k | 4.14 [3.12,5.16] | 41.72/13.5k | 1.45 [0.12,2.76] |
| Llama-3-style (1.23B) | 36.52/7.80k | 2.51 [1.78,3.24] | 41.28/3.79k | 3.61 [2.10,5.17] | 40.28/7.80k | 2.88 [2.08,3.69] |
| Pythia-style (1.08B) | 35.69/1.99k | 3.34 [1.81,4.87] | 39.55/5.58k | 5.34 [3.13,7.55] | 39.08/10.9k | 4.09 [1.79,6.38] |
| Gemma-3-style (1.0B) | 35.65/1.11k | 3.38 [1.47,5.29] | *42.44*/2.58k | 2.45 [1.08,3.82] | *42.39*/4.83k | 0.78 [0.03,1.52] |

Each Acc/FLOPs entry reports accuracy (%) followed by estimated training FLOPs (PF; k denotes $10^3$). Differences and confidence intervals are in percentage points. Boldface denotes TCT-SMIC, and italics denote the strongest baseline within each parameter group and ANLI round.

Given the layer input $Z^{(\ell)}$, the conditioned components are $X^{(\ell)}_{\text{real}} = \text{Re}\left(Z^{(\ell)}\right) + \xi^{(\ell)}_{\text{real}} \widehat{H}^{(\ell)}_{\text{real}}$, $X^{(\ell)}_{\text{imag}} = \text{Im}\left(Z^{(\ell)}\right) + \xi^{(\ell)}_{\text{imag}} \widehat{H}^{(\ell)}_{\text{imag}}$. Each $\xi^{(\ell)}_r$ is a zero-initialized learned scalar, so the conditioning contribution begins at zero.

### 3.3 Token–Context Attention under SMIC Conditioning

We project queries and values from the normalized real component, and project keys from the imaginary component: $\overline{X}^{(\ell)}_{\text{real}} = \text{LN}^{(\ell)}_1\left(X^{(\ell)}_{\text{real}}\right)$, $Q^{(\ell)}_a = \overline{X}^{(\ell)}_{\text{real}} W^{(\ell)}_{Q,a}$, $K^{(\ell)}_a = X^{(\ell)}_{\text{imag}} W^{(\ell)}_{K,a}$, $V^{(\ell)}_a = \overline{X}^{(\ell)}_{\text{real}} W^{(\ell)}_{V,a}$. Let $M^{(\ell)}_a$ denote the causal and key-padding mask for head $a$. The attention distribution is $A^{(\ell)}_a = \text{Softmax}\left(\frac{Q^{(\ell)}_a \left(K^{(\ell)}_a\right)^\top}{\sqrt{d_a}} + M^{(\ell)}_a\right)$. Rows corresponding to padded queries are zeroed. The multi-head output is then $O^{(\ell)} = \text{Concat}^H_{a=1}\left(\text{Dropout}\left(A^{(\ell)}_a\right) V^{(\ell)}_a\right) W^{(\ell)}_O$. After computing the multi-head output, the residual branch updates the real component as $U^{(\ell)} = X^{(\ell)}_{\text{real}} + \text{Dropout}\left(\text{LN}^{(\ell)}_2\left(O^{(\ell)}\right)\right)$. The feed-forward sublayer then processes the resulting representation: $F^{(\ell)} = \text{Dropout}\left(\text{ReLU}\left(U^{(\ell)} W^{(\ell)}_1 + b^{(\ell)}_1\right) W^{(\ell)}_2 + b^{(\ell)}_2\right)$. Adding this feed-forward contribution yields the next-layer real component, $T^{(\ell+1)} = U^{(\ell)} + F^{(\ell)}$. In contrast, the conditioned imaginary component is propagated without further transformation: $C^{(\ell+1)} = X^{(\ell)}_{\text{imag}}$. The two components are then recombined to form the next complex representation: $Z^{(\ell+1)} = T^{(\ell+1)} + iC^{(\ell+1)}$. Our current model configuration has three attention layers, 32 heads, and a feed-forward width of 4096. Classification uses the final-layer real component at the last valid hypothesis position.

## 4 Results

Our primary ANLI accuracy comparisons evaluate models trained from random initialization. For the parameter-comparable configurations in Table 1, we use FP32 arithmetic and deterministic CUDA and cuDNN kernels, with TF32 and cuDNN benchmarking disabled. Each configuration uses its

Table 2: ANLI-R1 comparison with encoders.

| Model | Params | Acc. (%) | Diff [95% CI] |
|---|---|---|---|
| TCT-SMIC | 0.80B | **39.03** | — |
| ModernBERT-large | 0.40B | 35.25 | 3.78 [1.44, 6.12] |
| DeBERTa-v2-xlarge | 0.89B | 33.30 | 5.73 [4.67, 6.79] |
| RoBERTa-large | 0.36B | 33.30 | 5.73 [4.67, 6.79] |

default tokenizer. We tune learning rate and dropout on ANLI-R1 using five seeds and select the configuration with the highest mean development accuracy. The selected hyperparameters are then reused for ANLI-R2 and ANLI-R3. Each ANLI round is trained and early-stopped separately with a patience of 5. We evaluate five additional seeds and report ten-seed mean accuracies with seed-paired 95% confidence intervals for pairwise differences. The from-scratch auxiliary analyses and ablations use the same seeds. The encoder comparison is restricted to ANLI-R1 and uses the same ten training seeds and hyperparameter tuning. We evaluate pretrained ModernBERT and Llama backbones with and without SMIC through deletion/insertion diagnostics. All reported runs use NVIDIA GH200 accelerators with eight CPU cores and 128 GB of host memory.

### 4.1 Performance comparisons on ANLI

Table 1 reports test accuracy on ANLI-R1, ANLI-R2, and ANLI-R3. TCT-SMIC achieves the highest mean accuracy among all configurations in Table 1 on each ANLI round.

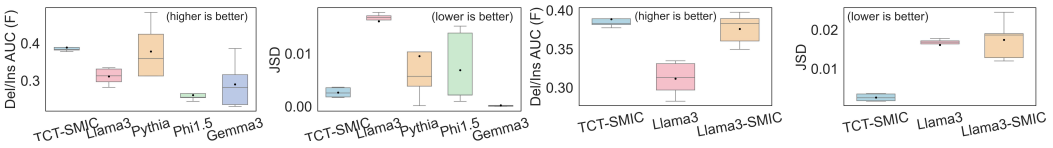
**Encoder comparisons.** Table 2 compares TCT-SMIC with randomly initialized ModernBERT-large, DeBERTa-v2-xlarge, and RoBERTa-large on ANLI-R1. TCT-SMIC achieves a mean accuracy of 39.03%, compared with 35.25% for ModernBERT-large. DeBERTa-v2-xlarge and RoBERTa-large each obtain 33.30% under the evaluated training settings. A related observation appears in Krishna et al. (2023), who report 31.82% MNLI validation accuracy for RoBERTa-base trained from random initialization. This provides context for the near-chance results, although it does not establish their cause in our experiments.

### 4.2 Attention-Based Diagnostics under Fixed-Length Token Perturbations

We use fixed-length token-ID replacements for the attention diagnostics. We sample 1000 labeled ANLI-R1 development examples with seed 42. For TCT-SMIC, we average results across checkpoints trained with the same seeds. Premises and hypotheses are tokenized separately with the Llama-3.2-1B tokenizer, with maximum lengths of 110 and 40, respectively. We replace token IDs after tokenization while preserving sequence length, attention masks, and token indices. The original and perturbed attention distributions therefore share the same token-index set.

For each $r \in \{1, 2, 4\}$, we generate three deterministic realizations per example and condition. The stopword-targeted condition samples eligible single-token stopwords, falling back to all eligible non-special positions when none are available. Replacement IDs are drawn from non-special entries that decode to alphabetic strings of at least two characters. If fewer than $r$ positions are available, all available positions are replaced. The random condition samples positions from all eligible non-special positions.

**Deletion and insertion.** Deletion and insertion are evaluated on the unmodified inputs. For each plotted configuration, we average final-layer attention across heads. Let $\mathbf{A}_i$ be the head-averaged attention matrix for example $i$, and let $L_i$ denote the query position used to extract the attention ranking for the evaluated configuration. For TCT-SMIC, $L_i$ is the final valid hypothesis position. The ranking score for candidate position $j$ is $s_{i,j} = A_{i,L_i,j}$. Let $\mathcal{D}_i$ be the set of non-special positions excluding $L_i$. We define $K_i = \lceil 0.1\,|\mathcal{D}_i| \rceil$. Let $p_{i,0}$ be the confidence of the original prediction. Replacing the top $t$ ranked positions with a fixed non-special placeholder yields a normalized confidence drop $d_{i,t} = \frac{\max\{0, p_{i,0} - p_{i,t}^{\mathrm{del}}\}}{p_{i,0}}$.




(a) Baseline AUC. (b) Baseline JSD. (c) Llama-3-style AUC. (d) Llama-3-style JSD.

Figure 2: Attention diagnostics on ANLI-R1.

For insertion, all positions in $\mathcal{D}_i$ are first replaced by the placeholder, after which the top $t$ ranked positions are restored: $r_{i,t} = \mathrm{clip}\left(\frac{p^{\mathrm{ins}}_{i,t}}{p_{i,0}}, 0, 1\right)$. Let $0 = t_{i,0} < t_{i,1} < \cdots < t_{i,M_i} = K_i$ denote the evaluated intervention steps, and let $x_{i,m} = \frac{t_{i,m}}{K_i}$. The deletion and insertion AUCs are $\mathrm{AUC}_{i,\mathrm{del}} = \sum_{m=1}^{M_i} (x_{i,m} - x_{i,m-1}) \frac{d_{i,t_{i,m-1}} + d_{i,t_{i,m}}}{2}$ and $\mathrm{AUC}_{i,\mathrm{ins}} = \sum_{m=1}^{M_i} (x_{i,m} - x_{i,m-1}) \frac{r_{i,t_{i,m-1}} + r_{i,t_{i,m}}}{2}$. The combined per-example diagnostic is $\text{DI-AUC}_{i,0.1} = \frac{1}{2}(\mathrm{AUC}_{i,\mathrm{del}} + \mathrm{AUC}_{i,\mathrm{ins}})$. The reported score is $\text{DI-AUC}_{0.1} = \frac{1}{N}\sum_{i=1}^{N} \text{DI-AUC}_{i,0.1}$.

**Conditional pooled-attention JSD.** For each example $i$, let $\Omega_i = \mathcal{D}_i$ be the common token-index set excluding special tokens and the final pooling token. Let $a_{i,j}$ and $a'_{i,j}$ denote the head-averaged original and perturbed pooled-query attention weights normalized over valid positions. The metric is defined when $|\Omega_i| \geq 2$. Restricting the weights to $\Omega_i$, clipping them below at $\epsilon = 10^{-8}$, and renormalizing yields distributions $\mathbf{p}_i$ and $\mathbf{q}_i$. For every $j \in \Omega_i$, $p_{i,j} = \frac{\max(a_{i,j},\epsilon)}{\sum_{v\in\Omega_i}\max(a_{i,v},\epsilon)}$, $q_{i,j} = \frac{\max(a'_{i,j},\epsilon)}{\sum_{v\in\Omega_i}\max(a'_{i,v},\epsilon)}$. Using natural logarithms, define $\mathbf{m}_i = \frac{\mathbf{p}_i + \mathbf{q}_i}{2}$, $J_i = \frac{1}{2}\mathrm{KL}(\mathbf{p}_i \,\|\, \mathbf{m}_i) + \frac{1}{2}\mathrm{KL}(\mathbf{q}_i \,\|\, \mathbf{m}_i)$. We report $\overline{J} = \frac{1}{N}\sum_{i=1}^{N} J_i$.

**Proposition 1** (Conditional Subset-Mass Bound)**.** *For any collection of subsets $\mathcal{S}_i \subseteq \Omega_i$,* $\frac{1}{N}\sum_{i=1}^{N} |p_i(\mathcal{S}_i) - q_i(\mathcal{S}_i)| \leq \min\left\{1, \sqrt{2\overline{J}}\right\}$.

The proof is provided in Appendix B. Proposition 1 applies directly to the top-10% prefix $\mathcal{S}_i^{0.1}$ formed by the $K_i$ highest-ranked positions under $\mathbf{p}_i$. This deterministic finite-sample inequality applies only to the restricted, smoothed, and renormalized distributions. It is neither a confidence bound nor a bound on the total full-row attention mass assigned to $\Omega_i$.

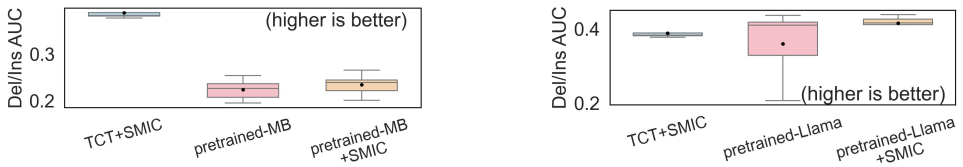
Deletion/insertion AUC tracks confidence under attention-ranked interventions, whereas conditional JSD measures redistribution within the normalized attention support. Figures 2a and 2b show that TCT-SMIC combines relatively high top-10% deletion/insertion AUC with low conditional JSD.

**SMIC as a plug-in for the Llama-3-style configuration.** We add a real-valued SMIC module to the Llama-3-style configuration to test its applicability beyond complex-valued TCT-SMIC. All other components and training settings are unchanged. Figure 2c shows that adding SMIC increases the deletion/insertion summary score, while Figure 2d reports the corresponding conditional pooled-attention JSD.

**SMIC in pretrained backbones.** We evaluate pretrained ModernBERT and Llama backbones with and without SMIC using the top-10% deletion/insertion diagnostic. Figures 3a and 3b show a higher median score with SMIC for both backbones. The increase is modest for ModernBERT, while Llama also exhibits a narrower interquartile range. These results show that the observed diagnostic effect is present in both pretrained encoder and decoder backbones and is not restricted to training from random initialization.

### 4.3 Empirical sensitivity under sparse lexical perturbations

The signed accuracy analysis uses pre-generated surface-text perturbations. This protocol differs from the fixed-length token-ID protocol used for the attention diagnostics in Section 4.2.




(a) Pretrained ModernBERT. (b) Pretrained Llama.

Figure 3: Top-10% deletion/insertion AUC on ANLI-R1 for pretrained models with and w/o SMIC.

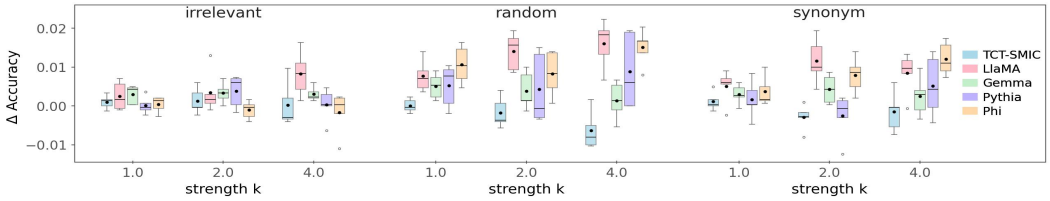



Figure 4: Signed accuracy change on ANLI-R1 under surface-text perturbations.

After removing unlabeled examples, we sample 1000 labeled examples from the ANLI-R1 development set with seed 42. The pre-generated data contain three perturbation families and replacement counts $k \in \{1, 2, 4\}$. Each value of $k$ has three perturbation realizations.

The SYNONYM family uses WordNet near-synonyms as a relatively conservative intervention, without assuming that each replacement preserves the original NLI label. The data label function-word and punctuation replacements as IRRELEVANT; we retain this name for consistency but do not treat it as a formal semantic-irrelevance judgment. The RANDOM family uses unconstrained lexical replacements and is treated as a corruption stress test.

The perturbations are generated once at the surface-text level, and the same perturbed premise–hypothesis strings are evaluated across models. Each configuration tokenizes the original and perturbed strings with its own tokenizer. Consequently, the same surface replacement may yield different subword segmentations across configurations.

For every model and condition, the reported signed accuracy change is $\Delta\,\mathrm{Acc} = \mathrm{Acc}_{\mathrm{original}} - \mathrm{Acc}_{\mathrm{perturbed}}$ . Positive values indicate lower accuracy after perturbation, whereas negative values indicate higher accuracy after perturbation.

**Finite-sample interpretation of signed accuracy change.** Fix one model configuration and one perturbation condition. Consider $N$ evaluated pairs $(x_i, x_i')$, where $x_i$ is the original input, $x_i'$ is its perturbed counterpart, and $y_i$ is the retained evaluation label. Let $\widehat{y}_i$ and $\widehat{y}_i'$ denote the corresponding predictions. Define $a_i = \mathbf{1}\{\widehat{y}_i = y_i\}$ , $a_i' = \mathbf{1}\{\widehat{y}_i' = y_i\}$ . Then $\Delta\,\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N}(a_i - a_i')$ .

Let $\mathcal{B} = \{i : a_i = 1, a_i' = 0\}$ denote predictions that become incorrect after perturbation, and let $\mathcal{G} = \{i : a_i = 0, a_i' = 1\}$ denote predictions that become correct. Then the signed accuracy change satisfies $\Delta\,\mathrm{Acc} = \frac{|\mathcal{B}|-|\mathcal{G}|}{N}$.

Let $\mathcal{W} = \{i : \widehat{y}_i \neq \widehat{y}_i', a_i = 0, a_i' = 0\}$ denote wrong-to-wrong prediction transitions. The prediction-flip rate is $\mathrm{FlipRate} = \frac{|\mathcal{B}|+|\mathcal{G}|+|\mathcal{W}|}{N}$. The signed accuracy change can therefore be written as $\Delta\,\mathrm{Acc} = \mathrm{BreakRate} - \mathrm{RecoveryRate}$, where $\mathrm{BreakRate} = \frac{|\mathcal{B}|}{N}$ and $\mathrm{RecoveryRate} = \frac{|\mathcal{G}|}{N}$. Thus, $\Delta\,\mathrm{Acc}$ is the break rate minus the recovery rate, whereas $\mathrm{FlipRate}$ also includes wrong-to-wrong transitions.

Signed accuracy change is not the proportion of changed predictions. Broken and recovered predictions can cancel, and transitions between two incorrect labels do not affect accuracy. Because the perturbed examples are not manually re-annotated, $\Delta\,\mathrm{Acc}$ can reflect both model sensitivity and changes in the correct NLI relation. We therefore interpret it as a protocol-specific empirical sensitivity measure rather than as label-preserving robust accuracy.

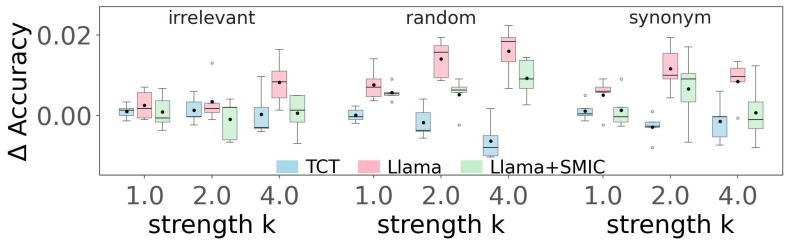



Figure 5: Signed accuracy change for TCT-SMIC and the Llama-3-style configurations with and without SMIC on the evaluated ANLI-R1 surface-text perturbations.

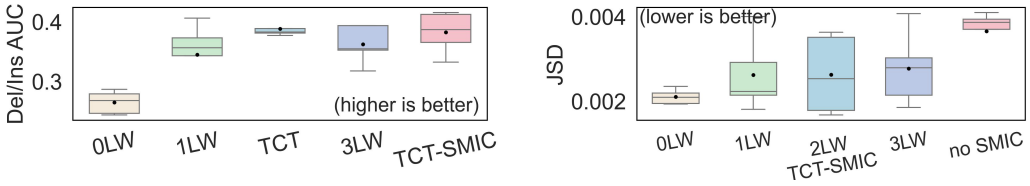



Figure 6: Attention diagnostics for SMIC ablations.

Figure 4 shows that TCT-SMIC has the smallest observed mean signed accuracy change across the evaluated ANLI-R1 conditions. For the synonym and function-word or punctuation families, this pattern corresponds to a smaller net accuracy change under the evaluated perturbation protocol. For random replacements, the result shows a smaller accuracy reduction under unconstrained corruption.

**SMIC as a plug-in for the Llama-3-style configuration.** Figure 5 shows that adding SMIC to the Llama-3-style configuration reduces the median $\Delta$Acc across the evaluated substitution families. This result shows that shared-reference conditioning reduces the median signed accuracy change under this protocol in the evaluated real-valued architecture.

### 4.4 Ablation

We test how SMIC design choices affect accuracy and perturbation sensitivity. We evaluate the number of learnable receptive-field scales, the removal of SMIC, the construction and cross-layer reuse of the shared input reference, and a layer-local reference variant. We report depth-matched and tokenizer-sensitivity checks.

**The number of learnable windows:** Starting from the default *2LW* setting, we vary the number of learnable token spans (*0LW/1LW/2LW/3LW*) across all *SMIC* layers, leaving other components unchanged. In particular, *0LW* disables the two Gaussian-mixture branches and retains the token-local and global branches.

On ANLI-R1, Fig. 6 indicates that the fixed-scale configuration produces the lowest attention JSD but also the lowest deletion/insertion summary score. The best validation checkpoint is reached 1.7 epochs later on average than with the default two-scale configuration across 10 seeds. Fixed scales therefore yield a distinct optimization and attention-diagnostic profile.

Figure 6 (left) shows that moving from *1LW* to *2LW* improves attention Del/Ins AUC, whereas *3LW* is not consistently better than *2LW* and has greater variability. Figure 6 (right) shows greater drift variability: *3LW* produces more volatile attention redistribution.

On ANLI-R3, Table 4 indicates that the two-scale configuration has the highest observed mean accuracy at 43.17%. The one-scale configuration reaches 42.47%, with a confidence interval for the difference that includes zero. The three-scale configuration reaches 41.57%, and its reported confidence interval favors the two-scale configuration. Among the evaluated settings, two learnable receptive-field scales give the best balance of mean accuracy and attention diagnostics. We do not claim that two scales are universally optimal outside this architecture and training setting.

Table 3: Mean ANLI accuracy with and without SMIC.

| Round | TCT-SMIC | w/o SMIC | Diff [95% CI] |
|---|---|---|---|
| R1 | 39.03 | 40.22 | $-1.19$ $[-2.67, 0.29]$ |
| R2 | 44.89 | 43.72 | $1.17$ $[0.07, 2.27]$ |
| R3 | 43.17 | 41.78 | $1.39$ $[0.62, 2.16]$ |

*Accuracy is reported in percent.

Table 4: Mean ANLI-R3 accuracy by number of learnable scales.

| Config | Acc | Diff [95% CI] |
|---|---|---|
| 2LW | 43.17 | — |
| 1LW | 42.47 | $0.70$ $[-0.57, 1.97]$ |
| 3LW | 41.57 | $1.60$ $[0.37, 2.83]$ |

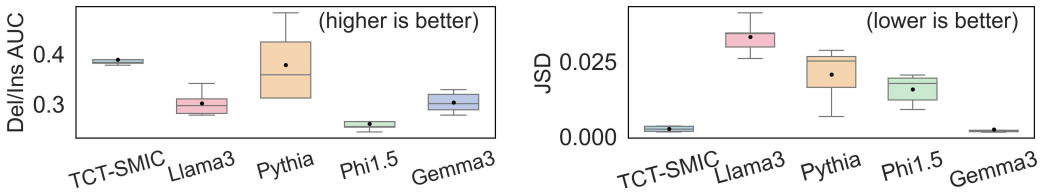



Figure 7: Attention diagnostics for three-layer models.

**Contribution of *SMIC*:** As Table 3 indicates, removing SMIC produces a round-dependent accuracy pattern. On ANLI-R1, the mean accuracy is 1.19 percentage points higher without SMIC, although the confidence interval includes zero. On ANLI-R2 and ANLI-R3, TCT-SMIC is higher by 1.17 and 1.39 percentage points, respectively. The attention diagnostics also show higher redistribution after SMIC is removed. Support for SMIC is strongest on ANLI-R2, ANLI-R3, and the perturbation diagnostics.

**Reference construction.** Replacing the multiscale construction of $\mathbf{Z}^{(0)}$ with a reference computed directly from token embeddings reduces mean ANLI-R2 accuracy from $44.89\%$ to $43.77\%$, a difference of $1.12$ percentage points ($95\%$ CI: $[0.17, 2.07]$). A layer-local variant that extracts token-span features from each layer output instead of the shared initial reference achieves $32.92\%$ on the same ANLI-R1 seeds, compared with $39.03\%$ for TCT-SMIC. These results support both multiscale reference construction and cross-layer reuse within the evaluated configurations.

**Depth-matched diagnostic:** We compare the three-layer TCT-SMIC with three-layer Llama-3-style, Phi-1.5-style, Pythia-style, and Gemma-3-style models under the same training protocol. Figure 7 shows that the higher deletion/insertion summary score and comparatively low attention redistribution persist in this depth-matched diagnostic. Thus, reduced depth alone does not explain the attention pattern.

**Tokenizer sensitivity check:** We rerun the ANLI-R1 seeds of TCT-SMIC with the Phi-1.5-style tokenizer. Mean accuracy is 39.03% with the Llama-3-style tokenizer and 39.54% with the Phi-1.5-style tokenizer. TCT-SMIC therefore retains comparable accuracy with a different tokenizer.

## 5 Limitations

First, the rule-generated surface-text perturbations are not manually re-annotated, and some, especially random replacements, may change the correct NLI relation. The resulting accuracy differences therefore measure protocol-specific empirical sensitivity rather than label-preserving robust accuracy. Second, because the baselines differ in architecture and tokenizer, the main comparison evaluates complete configurations rather than SMIC as an isolated variable. The within-architecture ablations and Llama-3-style plug-in provide more direct component-level evidence but do not cover all alternative conditioning mechanisms. Third, the main accuracy comparisons concern models trained from scratch for round-specific ANLI classification. The pretrained experiments cover two backbones and evaluate deletion/insertion AUC. Their scope does not establish broader task-performance gains, cross-round adversarial generalization, or applicability beyond NLI. Finally, deletion/insertion AUC and conditional pooled-attention JSD are descriptive diagnostics, not causal explanations or faithfulness guarantees.

## AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper and to support the retrieval and discovery of related work. References identified with the assistance of generative AI were independently verified against the original publications before being cited. All AI-assisted writing was reviewed and revised by the authors. The authors take full responsibility for the final content of this work, including its text, claims, citations, analyses, and artifacts.

## REFERENCES

Qianglong Chen, Feng Ji, Xiangji Zeng, Feng-Lin Li, Ji Zhang, Haiqing Chen, and Yin Zhang. Kace: Generating knowledge aware contrastive explanations for natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2516–2527, 2021.

Israel Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Transactions on speech and audio processing*, 11(5):466–475, 2003.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. Towards robustness against natural language word substitutions. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ks5nebunVn_.

Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341, 2015.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pp. 1243–1252. PMLR, 2017.

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI systems with sentences that require simple lexical inferences. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 650–655, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2103. URL https://aclanthology.org/P18-2103/.

Zhuoqun Huang, Neil G Marchant, Keane Lucas, Lujo Bauer, Olga Ohrimenko, and Benjamin Rubinstein. Rs-del: Edit distance robustness certificates for sequence classifiers via randomized deletion. *Advances in Neural Information Processing Systems*, 36:18676–18711, 2023.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4129–4142, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1423. URL https://aclanthology.org/D19-1423/.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Pride Kavumba, Ana Brassard, Benjamin Heinzerling, and Kentaro Inui. Prompting for explanations improves adversarial nli. is this true?{Yes} it is {true} because {it weakens superficial cues}. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2165–2180, 2023.

Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181/.

Kundan Krishna, Saurabh Garg, Jeffrey P Bigham, and Zachary C Lipton. Downstream datasets make surprisingly good pretraining corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12207–12222, 2023.

ChiYan Lee, Hideyuki Hasegawa, and Shangce Gao. Complex-valued neural networks: A comprehensive survey. *IEEE/CAA Journal of Automatica Sinica*, 9(8):1406–1426, 2022.

Rainer Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on speech and audio processing*, 9(5):504–512, 2001.

Zhao Meng, Yihan Dong, Mrinmaya Sachan, and Roger Wattenhofer. Self-supervised contrastive learning with adversarial perturbations for defending word substitution-based attacks. *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 87–101, 2022.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, 2020.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jackson Petty, Sjoerd Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. The impact of depth on compositional generalization in transformer language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7239–7252, 2024.

Zhen Tian, Wayne Xin Zhao, Changwang Zhang, Xin Zhao, Zhongrui Ma, and Ji-Rong Wen. Eulerformer: Sequential user behavior modeling with complex vector attention. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1619–1628, 2024.

Jie Wang, XiaoGuang Hu, Enyi Guan, Tongyun Li, ZhuShun Ding, and Yidong Yao. A novel feed-forward segmented digital automatic gain control algorithm for long-term evolution digital radio-over-fibre systems. *Iet Communications*, 11(15):2362–2369, 2017.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. In *ICLR Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023a. Available at https://openreview.net/forum?id=uw6HSkgoM29.

Wenjie Wang, Pengfei Tang, Jian Lou, and Li Xiong. Certified robustness to word substitution attack with differential privacy. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1102–1112, 2021a.

Xiaosen Wang, Jin Hao, Yichen Yang, and Kun He. Natural language adversarial defense through synonym encoding. In *Uncertainty in Artificial Intelligence*, pp. 823–833. PMLR, 2021b.

Zhaoyang Wang, Zhiyue Liu, Xiaopeng Zheng, Qinliang Su, and Jiahai Wang. Rmlm: A flexible defense framework for proactively mitigating word-level adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2757–2774, 2023b.

Yichen Yang, Xin Liu, and Kun He. Synonym-unaware fast adversarial training against textual adversarial attacks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 727–739, 2025.

Xingcheng Yao, Yanan Zheng, Xiaocong Yang, and Zhilin Yang. Nlp from scratch without large-scale pretraining: A simple and efficient framework. In *International Conference on Machine Learning*, pp. 25438–25451. PMLR, 2022.

Muchao Ye, Ziyi Yin, Tianrong Zhang, Tianyu Du, Jinghui Chen, Ting Wang, and Fenglong Ma. Unit: a unified look at certified robust training against text adversarial perturbation. *Advances in Neural Information Processing Systems*, 36:22351–22368, 2023.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. Certified robustness to text adversarial attacks by randomized [mask]. *Computational Linguistics*, 49(2):395–427, 2023.

Xin Zhang and Victor S. Sheng. Token-context attention for NLI: An alternative to self-attention. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)*, 2026. URL https://openreview.net/forum?id=sCselFSf60.

Xinyu Zhang, Hanbin Hong, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. Text-crs: A generalized certified robustness framework against textual adversarial attacks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 2920–2938. IEEE, 2024a.

Zeliang Zhang, Wei Yao, Susan Liang, and Chenliang Xu. Random smooth-based certified defense against text adversarial attack. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1251–1265, 2024b.

Haiteng Zhao, Chang Ma, Xinshuai Dong, Anh Tuan Luu, Zhi-Hong Deng, and Hanwang Zhang. Certified robustness against natural language attacks by causal intervention. In *International Conference on Machine Learning*, pp. 26958–26970. PMLR, 2022.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. Defense against synonym substitution-based adversarial attacks via Dirichlet neighborhood ensemble. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5482–5492, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.426. URL https://aclanthology.org/2021.acl-long.426/.

# A Derivation of the cartesian-form of the complex-vector token representation

Fix a token index $j \in \{1, \dots, n\}$ and an embedding coordinate $k \in \{1, \dots, d\}$. Define $G_k = \sqrt{\sum_{i=1}^{n} w_{i,k}^2}$. If $G_k = 0$, then $w_{i,k} = 0$ for every token position $i$. In this case, the numerically stabilized implementation returns $Z_{j,k} = 0$. The exact derivation below therefore considers the nondegenerate case $G_k > 0$. Define the adapted magnitude $G'_{j,k} = G_k + w_{j,k}$, the normalized coordinate $c_{j,k} = \frac{w_{j,k}}{G_k}$, and the phase $\alpha_{j,k} = \operatorname{atan2}\left(\sqrt{1 - c_{j,k}^2}, c_{j,k}\right)$.

In the implementation, denominators are lower-bounded by $\epsilon = 10^{-8}$, squared cosine values are clipped to $[0, 1]$, and the square-root argument is lower-bounded by $\epsilon$. The following equalities describe the corresponding idealized expression when these numerical safeguards are inactive.

Define the *Complex-Vector Token Representation (CVTR)* in polar form as $Z_{j,k} := G'_{j,k} e^{i\alpha_{j,k}}$. Using Euler's formula $e^{i\alpha} = \cos(\alpha) + i \sin(\alpha)$, we obtain $Z_{j,k} = G'_{j,k}\big(\cos(\alpha_{j,k}) + i \sin(\alpha_{j,k})\big) = T_{j,k} + iC_{j,k}$.

**Real part:** $T_{j,k} = \Re(Z_{j,k}) = (G_k + w_{j,k}) \cos(\alpha_{j,k}) = (G_k + w_{j,k}) \frac{w_{j,k}}{G_k} = w_{j,k} + \frac{w_{j,k}^2}{G_k} = w_{j,k} \left(1 + \frac{w_{j,k}}{G_k}\right) = (1 + \cos(\alpha_{j,k}))\, w_{j,k}$.

**Imaginary part:** $C_{j,k} = \Im(Z_{j,k}) = (G_k + w_{j,k})\sin(\alpha_{j,k}) = (G_k + w_{j,k})\frac{\sqrt{G_k^2 - w_{j,k}^2}}{G_k} = \left(1 + \frac{w_{j,k}}{G_k}\right)\sqrt{G_k^2 - w_{j,k}^2} = (1 + \cos(\alpha_{j,k}))\sqrt{G_k^2 - w_{j,k}^2}$. Finally, since $G_k^2 = \sum_{i=1}^n w_{i,k}^2$, we have $G_k^2 - w_{j,k}^2 = \sum_{i \neq j} w_{i,k}^2$, and therefore $C_{j,k} = (1 + \cos(\alpha_{j,k}))\sqrt{\sum_{i \neq j} w_{i,k}^2}$.

Thus, the deterministic factor $1 + \cos(\alpha_{j,k})$ rescales both Cartesian components under the adapted magnitude $G'_{j,k} = G_k + w_{j,k}$. It is not an independently parameterized gating module.

## B Proof of Proposition 1

*Proof.* For two probability distributions $\mathbf{p}$ and $\mathbf{q}$ on the same finite support, define their total variation distance by $\mathrm{TV}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1$.

For every subset $\mathcal{S}$ of the common support, $|p(\mathcal{S}) - q(\mathcal{S})| \leq \mathrm{TV}(\mathbf{p}, \mathbf{q})$. Let $\mathbf{m} = \frac{\mathbf{p}+\mathbf{q}}{2}$.

Pinsker's inequality gives $\mathrm{KL}(\mathbf{p}\|\mathbf{m}) \geq 2\,\mathrm{TV}(\mathbf{p}, \mathbf{m})^2$ and $\mathrm{KL}(\mathbf{q}\|\mathbf{m}) \geq 2\,\mathrm{TV}(\mathbf{q}, \mathbf{m})^2$. Because $\mathrm{TV}(\mathbf{p}, \mathbf{m}) = \mathrm{TV}(\mathbf{q}, \mathbf{m}) = \frac{1}{2}\mathrm{TV}(\mathbf{p}, \mathbf{q})$, the Jensen–Shannon divergence satisfies $J(\mathbf{p}, \mathbf{q}) \geq \frac{1}{2}\mathrm{TV}(\mathbf{p}, \mathbf{q})^2$. It follows that $|p(\mathcal{S}) - q(\mathcal{S})| \leq \sqrt{2J(\mathbf{p}, \mathbf{q})}$.

Applying this inequality to every evaluated pair gives $\frac{1}{N}\sum_{i=1}^N |p_i(\mathcal{S}_i) - q_i(\mathcal{S}_i)| \leq \frac{1}{N}\sum_{i=1}^N \sqrt{2J_i}$. Since the square-root function is concave, Jensen's inequality gives $\frac{1}{N}\sum_{i=1}^N \sqrt{2J_i} \leq \sqrt{\frac{2}{N}\sum_{i=1}^N J_i} = \sqrt{2\overline{J}}$. Since $p_i(\mathcal{S}_i)$ and $q_i(\mathcal{S}_i)$ both lie in $[0, 1]$, we also have $|p_i(\mathcal{S}_i) - q_i(\mathcal{S}_i)| \leq 1$ for every $i$. Averaging over the evaluated pairs gives $\frac{1}{N}\sum_{i=1}^N |p_i(\mathcal{S}_i) - q_i(\mathcal{S}_i)| \leq 1$. Combining this bound with the Jensen–Shannon bound yields $\frac{1}{N}\sum_{i=1}^N |p_i(\mathcal{S}_i) - q_i(\mathcal{S}_i)| \leq \min\left\{1, \sqrt{2\overline{J}}\right\}$, which proves the proposition. □

*Remark* 1. If Jensen–Shannon divergence is computed with base-2 logarithms, the right-hand side of Proposition 1 becomes $\min\left\{1, \sqrt{2\ln(2)\overline{J}}\right\}$.