Token-Wise Kernels (TWiKers) for Vicinity-Aware Attention in Transformers

Anonymous ACL submission

Abstract

Self-attention mechanisms in transformers enable tokens to interact across a sequence but lack an explicit inductive bias to capture local contextual dependencies, an inherent characteristic of human languages. We propose Token-Wise Kernels (`TWiKers`), a novel enhancement to transformers that learn token-specific convolutional kernels applied to the keys or values. Each token is assigned a small kernel, initialized to the "Central Dirac" (e.g., [0,1,0] for size=3), meaning the token "bears" the attention from all other tokens alone. During training, these kernels adapt, and greater deviation from the Central Dirac indicates stronger attention redistribution to neighboring tokens. This introduces the first transformer weights with direct semantic interpretability. Our experiments show that content words (e.g., nouns and verbs) retain self-focus, while function words (e.g., prepositions and conjunctions) shift attention toward their neighbors, aligning with their syntactic and semantic roles. We further apply `TWiKers` to distinguish literary genres, historical periods, and authors, demonstrating their effectiveness in capturing high-level stylistic patterns. Finally, we demonstrate the potential of `TWiKers` as an effective inductive bias to improve transformer training, validated across a range of downstream tasks.

1 Introduction

Transformers have revolutionized natural language processing (NLP), powering large language models (LLMs) that achieve state-of-the-art performance across diverse tasks. Recent base models, such as DeepSeek-V3 (DeepSeek-AI et al., 2025), LLaMA-4 (Grattafiori et al., 2024), and Qwen-3 (Yang et al., 2025), have exhibited increasingly strong emergent abilities, fueling speculation that large language models may be approaching the threshold of artificial general intelligence (AGI).

One of the most remarkable aspects of transformers is the multi-head attention mechanism (Vaswani et al., 2017), which not only offers scalability but also enhances interpretability. Deep embeddings facilitate distance-based comparisons, a fundamental principle behind retrieval-augmented generation (RAG) (Lewis et al., 2020)–a key ingredient of modern AI agents. Token (shallow) embeddings are also widely used for lexical analysis, including clustering (Cha et al., 2017; Zhang et al., 2023), visualization (Le and Lauw, 2017; Molino et al., 2019), and analogy reasoning (Zhu et al., 2018; Petersen and van der Plas, 2023). However, these embeddings lack inherent meaning on their own; their interpretability depends on distance measurements and comparisons. So far, *no weights in transformers have been shown to encode direct semantic meaning at the parameter level.*

While one strength of transformers is their ability to capture long-range contextual dependencies, human languages exhibit strong vicinity reliance, particularly at the lexical level. For example, when reading *"War and Peace"*, a human would naturally focus on *"War"* and *"Peace"* while ignoring *"and"*, which carries less semantic weight. This selective attention to content words over function words is a fundamental characteristic of natural language, not unique to English but observed in most languages. Such locality has supported sliding-window attention, enabling models like Longformer (Beltagy et al., 2020) to achieve linear-time attention computation, along with its variations such as BigBird (Zaheer et al., 2020), Mamba (Gu and Dao, 2024), and LongLoRA (Chen et al., 2024). In computer vision, similar principles have been applied in models like Swin Transformer (Liu et al., 2021) and Neighborhood Attention Transformer (NAT) (Hassani et al., 2022). Another approach that exploits local dependencies is n-gram tokenization, which explicitly captures fixed-length word sequences (Mikolov et al., 2013b; Pennington et al., 2014; Bojanowski et al., 2017; Devlin et al., 2019). However, *despite the prevalence of local dependencies in hu-*

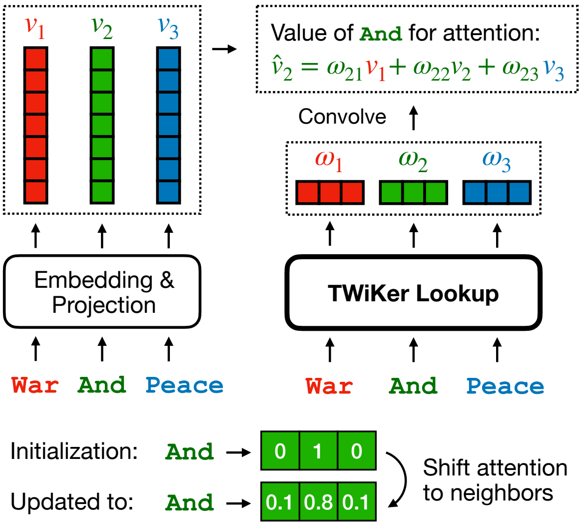


Figure 1: Overview of the `TWiKer` mechanism. After training, ω deviating from the Central Dirac ([0,1,0]) indicates a shift in attention toward neighboring tokens. Here we omit `TWiKers` for keys and their variability across heads for simplicity.

man languages, the transformer architecture lacks an explicit inductive bias to take advantage of this characteristic.

In this paper, we introduce **Token-Wise Kernels (`TWiKers`)**, a novel enhancement to transformers that incorporates an inductive bias to reflect vicinity reliance while preserving transformer's global attention. We assign a small, trainable convolutional kernel to each token, enabling the model to learn how different tokens interact with their immediate neighbors through attention redistribution. In this way, `TWiKers` capture vicinity-aware semantic relationships, as illustrated in Figure 1.

The key novelties of `TWiKers` are as follows:

1. **Direct Semantic Meaning**: Unlike standard transformer weights, `TWiKers` learn interpretable patterns that align with syntactic and semantic roles of words. For example, content words (nouns, verbs) tend to retain self-focus, while function words (e.g., prepositions, conjunctions) emphasize their surroundings.

2. **Automatic Lexical and Semantic Analysis**: Since `TWiKers` encode token-specific contextual behavior, they can be directly analyzed to distinguish lexical categories, track historical language changes, and classify text styles without additional supervision.

3. **Enhanced Training Efficiency**: Given its semantic relevance, `TWiKers` provide a meaningful inductive bias that may improve both pretraining and finetuning by helping transformers learn embeddings aligning better with human languages.

We validate `TWiKers` through comprehensive experiments for English, demonstrating their alignment with linguistic principles and their effectiveness in real-world applications.

2 Related Work

2.1 Sliding-Window Attention

To address the quadratic complexity of full self-attention, the sliding-window methods confine attention to local regions. For example, Longformer (Beltagy et al., 2020) uses fixed-size local windows with select global tokens for linear complexity, while BigBird (Zaheer et al., 2020) integrates random and sparse global patterns to better approximate full attention. Recent methods like Mamba (Gu and Dao, 2024), LongLoRA (Chen et al., 2024), BASED (Arora et al., 2024), and CEPE (Yen et al., 2024) further optimize local attention. In computer vision, approaches such as Swin Transformer (Liu et al., 2021) and NAT (Hassani et al., 2022) similarly enhance efficiency by focusing attention on local regions.

Although sliding-window approaches resemble `TWiKers` in their emphasis on local context, their motivations and effects fundamentally differ. Sliding-window methods aim to improve efficiency by restricting attention to fixed-size windows, thereby compromising the transformer's global receptive field. In contrast, `TWiKers` explicitly encode local semantic interactions into token-level parameters, enabling the model to capture local dependencies without sacrificing global attention. Nonetheless, both approaches are grounded in the vicinity-dominated nature of human languages.

2.2 N-Gram Tokenization

N-gram tokenization, also based on strong vicinity reliance, represents language as sequences of contiguous units. Traditional n-gram models–often enhanced by smoothing techniques such as Kneser-Ney (Kneser and Ney, 1995)–have demonstrated effectiveness in classical language modeling. Neural approaches further incorporate n-gram features: fastText (Bojanowski et al., 2017) enriches word embeddings with character-level n-grams, while BPE (Sennrich et al., 2016) and SentencePiece (Kudo and Richardson, 2018) construct subword vocabularies based on frequent n-gram patterns. Recent developments have extended the power of n-gram modeling. N-Grammer (Thai et al., 2020) augments transformers by integrating latent n-gram representations directly into the architecture. Subsequent analytical work has employed n-gram statistics to examine how language models implicitly capture linguistic structures (Li et al., 2022), conceptually close to our methodology. The Infini-gram model (Liu et al., 2024) generalizes n-gram methods to infinite-length sequences using an advanced back-off mechanism. Again, n-gram tokenizers highlight the strong local dependencies in natural language, which modern subword tokenizers under-exploit. This principle aligns with our approach. However, `TWiKers` capture locality through adaptive, semantically meaningful weights learned directly from data.

2.3 Token Embeddings in NLP Tasks

Token embeddings are shallow representations of tokens. While they are less effective than deep transformer embeddings for contextual understanding, they have proven valuable in lexical semantic studies. Foundational models such as LSA (Landauer and Dumais, 1997), word2vec (Mikolov et al., 2013b), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017) laid the groundwork for applications including clustering (Hill et al., 2015; Vulić and Mrkšić, 2018), visualization (Mikolov et al., 2013a; Reif et al., 2019), and analogy reasoning (Mikolov et al., 2013b). Recent work has extended these embeddings to cognitive and psycholinguistic domains, where they are used to model human semantic memory, word associations, and lexical access (Günther et al., 2019; Nematzadeh et al., 2017; Chronis and Erk, 2020; Samir et al., 2020). However, existing token embeddings are largely derived from statistical co-occurrence and offer limited semantic interpretability via distance comparison. In contrast, `TWiKers` provide direct semantic interpretability, distinguishing lexical categories (e.g., content vs. function words) and enabling automatic, linguistically meaningful analysis without supervision.

3 Methodology

3.1 Token-Wise Kernels in Self-Attention

In a standard transformer architecture (Vaswani et al., 2017), the attention mechanism computes output representations using the scaled dot-product attention, enabling each token to attend globally. To introduce an explicit inductive bias for vicinity awareness while preserving global dependencies, we associate each token in the vocabulary with two kernels of size n: a key kernel $\boldsymbol{\omega}^{\mathrm{k}} \in \mathbb{R}^n$ and a value kernel $\boldsymbol{\omega}^{\mathrm{v}} \in \mathbb{R}^n$. These kernels modify the attention mechanism by convolving the keys and values with the kernels:

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}(\boldsymbol{\Omega}^{\mathrm{k}}\boldsymbol{K})^\top}{\sqrt{d}}\right)(\boldsymbol{\Omega}^{\mathrm{v}}\boldsymbol{V}), \quad (1)$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{A} \in \mathbb{R}^{L\times d}$ are the matrices of query, key, value, and attention score (L: sequence length; d: feature dimension; batch and head dimensions omitted), and $\boldsymbol{\Omega}^{\mathrm{k}}, \boldsymbol{\Omega}^{\mathrm{v}} \in \mathbb{R}^{L\times L}$ are banded matrices with bandwidth n, assembled from the per-token kernels ω^{k}_{ij} and ω^{v}_{ij} ($i = 1,\dots,L$; $j = 1,\dots,n$) in a sliding-window manner. For example, when $L = 4$ and $n = 3$:

$$\boldsymbol{\Omega}^{\mathrm{k}} = \cancel{\omega^{\mathrm{k}}_{11}}\begin{bmatrix} \omega^{\mathrm{k}}_{12} & \omega^{\mathrm{k}}_{13} & & \\ \omega^{\mathrm{k}}_{21} & \omega^{\mathrm{k}}_{22} & \omega^{\mathrm{k}}_{23} & \\ & \omega^{\mathrm{k}}_{31} & \omega^{\mathrm{k}}_{32} & \omega^{\mathrm{k}}_{33} \\ & & \omega^{\mathrm{k}}_{41} & \omega^{\mathrm{k}}_{42} \end{bmatrix}\cancel{\omega^{\mathrm{k}}_{43}}. \quad (2)$$

Here, ω^{k}_{11} and ω^{k}_{43} are omitted at the sequence boundaries to avoid padding. The value transformation matrix $\boldsymbol{\Omega}^{\mathrm{v}}$ is constructed analogously. For clarity, we present Equation (1) explicitly; in practice, we use a standard fold-multiply-unfold pipeline to maintain $\mathcal{O}(L)$ computational complexity.

Understanding the semantic significance of key convolution ($\boldsymbol{\Omega}^{\mathrm{k}}\boldsymbol{K}$) and value convolution ($\boldsymbol{\Omega}^{\mathrm{k}}\boldsymbol{V}$) is essential for interpreting the learned weights. Key convolution directly shifts attention by incorporating key representations from neighboring tokens, effectively redistributing focus to local context. Value convolution blends surrounding context into the retrieved representations, allowing each token to reflect nuanced semantic information from its vicinity. Together, these mechanisms enhance the model's ability to encode syntactic relationships and contextual meaning by explicitly reinforcing local dependencies. Notably, these vicinity-aware behaviors are semantically meaningful *only because the kernels are token-specific* rather than position-based, distinguishing `TWiKers` from position-wise parameterizations such as $(\mathrm{IA})^3$ (Liu et al., 2022).

3.2 Enforcing Causality

In autoregressive language modeling, tokens are not allowed to attend to future tokens (Vaswani et al., 2017; Dai et al., 2019). However, Eq. (1) introduces information leakage as `TWiKers` allow the i-th token to aggregate key representations from up to $(n-1)/2$ future tokens, thus violating causality. To address this, we restrict the range of key and value summation in Eq. (1). Specifically, the attention weights, $\boldsymbol{A} = \boldsymbol{Q}(\boldsymbol{\Omega}^{\mathrm{k}}\boldsymbol{K})^{\top}$, are revised as:

$$A_{ij} = \sum_{l=1}^{d} Q_{il} \sum_{m=1}^{\min(n,p+i-j)} \omega_{jm}^{\mathrm{k}} K_{j-p+m,l} \quad (3)$$

where $p = (n+1)/2$. The inner summation is now limited by $\min(n, p+i-j)$, ensuring that query i only attends to past and present keys. In practice, only the main diagonal and the first $p-2$ sub-diagonals of A_{ij} require correction, preserving$\mathcal{O}(L)$ complexity, incurring minimal overhead. The attention output is similarly adjusted by restricting the value aggregation range.

3.3 Enforcing Probabilistic TWiKers

To enhance the interpretability of `TWiKers`, we enforce them to be probabilistic distributions (non-negative and summing to one). We define the unconstrained trainable parameters $\hat{\boldsymbol{\omega}}^{\mathrm{k}}$ and $\hat{\boldsymbol{\omega}}^{\mathrm{v}}$, which are transformed via a softmax function to compute the actual kernels used for convolution:

$$\omega_{ij}^{\mathrm{k,v}} = \frac{\exp\left(\hat{\omega}_{ij}^{\mathrm{k,v}}/\tau\right)}{\sum_{m=1}^{n} \exp\left(\hat{\omega}_{mj}^{\mathrm{k,v}}/\tau\right)}, \quad i = 1, 2, \ldots, n, \quad (4)$$

where τ is the temperature hyperparameter.

To ensure that `TWiKers` do not affect the model prior to training, we initialize the unconstrained kernels to a sharpened Central Dirac, such as $[0, 10, 0]$ for $n = 3$. This initialization enforces self-focus at the beginning, allowing the model to learn meaningful vicinity-aware modifications during training.

4 Language-focused Experiments

In this section, we finetune GPT-2 (Radford et al., 2019) for causal language modeling on a range of English corpora spanning poetry, novels, drama, translations, and scientific articles, as summarized in Table 1. Although `TWiKers` are generally applicable to other languages and newer architectures, we focus on English and GPT-2 due to resource constraints (see Limitations). Data declarations and engineering details are provided in Appendices A, and B. To enhance comparability between corpora, we adopt the following setup:

Data sampling From each corpus, we sample 2200 segments containing complete sentences (up to 1000 tokens each); 2000 are used for training, 200 for evaluation.

Two-stage finetuning We finetune GPT-2 on each corpus independently. We observe that different corpora converge at different rates when trained with `TWiKers` (e.g., `HarryPotter` converges faster than `Shakespeare`). This discrepancy arises likely because each corpus starts at a different distance from the pretrained model's local minimum. To account for this, we first finetune each corpus for 30 epochs without `TWiKers`, then continue for 30 epochs with `TWiKers` enabled.

`TWiKer` configuration `TWiKers` applied to keys or values can both shift attention toward neighboring tokens. To enhance semantic interpretability, we do not activate `TWiKers` for keys and values at the same time. By default, we apply value convolution only, as it demonstrates greater robustness. The kernel size is fixed at three and shared across all attention heads. The softmax temperature is 0.4, and learning rates are fixed at 5×10^{-5} for model weights and 5×10^{-3} for `TWiKer` parameters, the latter compensating for small gradients near the Central Dirac initialization. In ablation studies, we compare different configurations (see Section 4.4 for details).

4.1 Lexical Attention Patterns

`TWiKers` provide direct insight into local attention behavior. Here, we analyze the learned `TWiKer` weights from the `HarryPotter` corpus. Figure 2 shows that content words (e.g., `"Potter"`, `"gold"`) have sharply peaked central weights, focusing attention on themselves, while function words (e.g., `"the"`, `"and"`) distribute attention over neighbors, reflecting their syntactic role in structuring phrases rather than anchoring meaning. This difference aligns well with traditional linguistic distinctions between semantic and grammatical categories.

To quantify this, we compute the average deviation from the Central Dirac for common parts of speech (PoS) tags. As shown in Figure 3, function words (e.g., determiners, conjunctions) show greater deviation, whereas content-rich categories (e.g., nouns, verbs) stay closer to the central peak. These results demonstrate that `TWiKers` can capture meaningful linguistic structure in an interpretable, unsupervised manner.

Corpus (Time Period) Data Source	Linguistic Characteristics
Shakespeare (1590–1616) Zahid (2021)	**Shakespeare's plays**: 17 plays with poetic diction, inverted syntax, metaphor, and rhetorical patterning, unlike modern English.
Victorian (1800–1900) Chapman (2022)	**British Poetry from the Victorian Era**: 2216 poems. Characterized by formal modifiers, measured syntax, and dense semantics.
NewPoems (post 2000) Poetry Foundation (2023)	**Contemporary Poetry**: 5000 poems. Free verse, irregular syntax, and playful imagery, reflecting creative and less constrained modern poetic style.
War&Peace (~1923) McKay (2016)	**English Translation of *War and Peace***: Formal style with Russian-influenced syntax, frequent passive constructions, and complex, multi-clause sentences.
RedChamber (~1979) Internet Archive (2020)	**English Translation of *The Dream of the Red Chamber***: Five translation versions included, exhibiting diverse stylistic choices while consistently preserving the classical Chinese narrative style.
Dickens (1836–1870) McAdams (2020)	**Novels by Charles Dickens**: 15 novels. Ornate prose with complex noun phrases and descriptive clauses; language shifts with character voice.
StKing (1980–2000) Ajmain (2022)	**Novels by Stephen King**: 20 novels. Direct language with active verbs and informal phrasing; blends colloquial realism with psychological tension.
HarryPotter (1997–2007) Kapoor (2024)	***Harry Potter***: All 7 novels. Clear, child-friendly prose with simple structures and verbs; combines fantasy world-building with British idioms.
Papers (post 2000) Holbrook (2020)	**Scientific Articles**: 1000 paragraphs. Information-dense, impersonal prose with nominalizations, passive voice, and technical terms.

Table 1: Corpora used for experiments, spanning diverse genres, time periods, and writing styles.

Lexical handedness We observe a directional asymmetry in learned TWiKer weights. For tokens whose central kernel weight is below 0.99, we categorize them as *left-handed* if the left value exceeds the right, and *right-handed* otherwise. In the HarryPotter corpus, 9,570 tokens are right-handed while only 84 are left-handed–a striking imbalance. This reflects the right-branching nature of English (Dryer, 1992; Du et al., 2020), where syntactic dependents such as complements and modifiers typically follow their heads. Function words (e.g., prepositions, subordinating conjunctions) often anticipate or introduce material to their right, naturally shifting attention forward in the sequence. Thus, TWiKers internalize not only lexical category behavior but also broader structural tendencies.

4.2 Cross-Corpus Comparison

Figure 4 summarizes TWiKer deviations from the Central Dirac across corpora, revealing four categories: Academic, Poetic, Translations, and Novels. These results demonstrate the strong influence of genre and stylistic conventions on attention patterns: more structured or constrained texts show lower deviations, while narrative-driven texts exhibit higher deviation.

The Academic corpus (Papers) exhibits the lowest deviation, reflecting rigid syntax and dense semantics that limit contextual dependencies and maintain tight lexical focus. Poetic corpora also show low deviation, reflecting their structured phrasing and rhythmic regularity–Victorian poetry has lower deviation than NewPoems: the former follows metrical and formal constraints, while the latter–including free verse and children's poetry–features more flexible syntax and modern phrasing, leading to greater attention spread. Shakespeare falls in between, combining poetic formality with syntactic inversion and dramatic rhythm.

Novels, both translated and native, display the highest deviations, reflecting their narrative characteristic and syntactic variety. However, translated works (War&Peace, originally written in 19th-century Russian, and RedChamber, from 18th-century vernacular Chinese) exhibit lower deviation than native English novels, likely reflecting the relative syntactic compactness of their source languages and regularization introduced during translation. Within novels, HarryPotter exhibits the highest deviation, reflecting its narrative style and flexible structures.

For more in-depth analysis, we examine TWiKer deviations across PoS tags in Appendix C, focusing

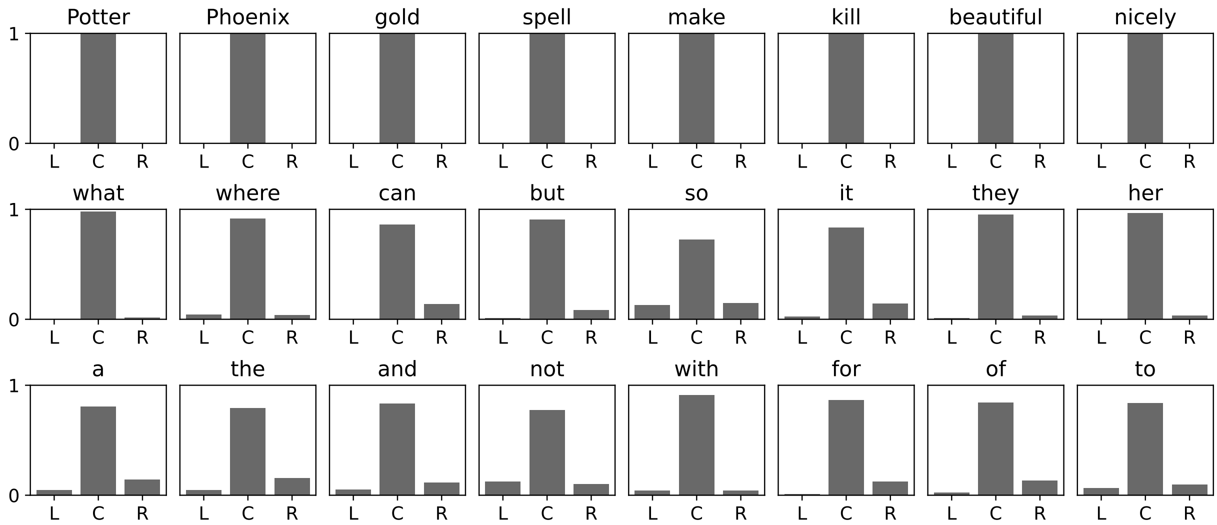


Figure 2: Learned `TWiKer` kernels for selected tokens in `HarryPotter`. Each triplet of bars shows the kernel weights for left (L), center (C), and right (R) positions. Content words show dominant center weights, while function words spread their attention to adjacent tokens.

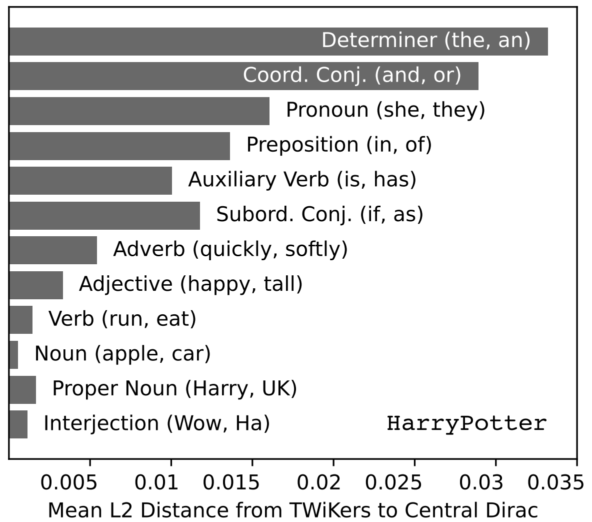


Figure 3: Mean deviation of learned `TWiKers` from Central Dirac $[0, 1, 0]$ across PoS tags in `HarryPotter`. Higher values indicate broader attention spread away from the token itself.

on three corpora that exhibit notable divergence from general English patterns.

4.3 Clustering Translations

As a real-world application, we use `TWiKers` to cluster different English translations of *The Dream of the Red Chamber* (红楼梦), one of the most celebrated novels in Chinese literature. A cloud over the novel's history is the uncertainty of its authorship. It is established that Cao Xueqin (曹雪芹) wrote the first 80 chapters, whereas the authorship of the final 40 chapters–possibly by Gao E (高鹗)–remains debated. While we are unable to resolve this historical mystery using GPT-2, it inspires us to analyze five full English translations of the novel through the lens of `TWiKers`.

We compare five English versions of Dream of the Red Chamber. The earliest, by H. Bencraft Joly (Joly, 1893), covers Chapters 1-56 in formal, archaic Victorian prose. It was later extended to Chapter 80 by Florence and Isabel McHugh (McHugh and McHugh, 1958), based not on the Chinese original but on Franz Kuhn's German version, adding an extra interpretive layer. The widely circulated edition by Yang Hsien-yi and Gladys Yang (Yang and Yang, 1980), published in China, is clear and faithful, prioritizing literal accuracy and accessibility over literary embellishment. David Hawkes' acclaimed translation (Hawkes and Minford, 1986) (Volumes I-III), completed by John Minford (Volumes IV-V), is widely accepted as the most literary version, with idiomatic prose and extensive cultural notes. Lastly, we include a machine-translated version by OpenAI's `o3-mini`, which is fluent and modern but may lack consistency in style between chapters.

For analysis, we split the novel's 120 chapters into five segments of roughly 24 chapters each, and treat each segment as a separate corpus for training `TWiKer`-enhanced GPT-2. Figure 5 shows the mean `TWiKer` deviation from the Central Dirac, indicating that even this single scalar metric can broadly distinguish between translation styles.

For finer-grained analysis, we compute the aver-

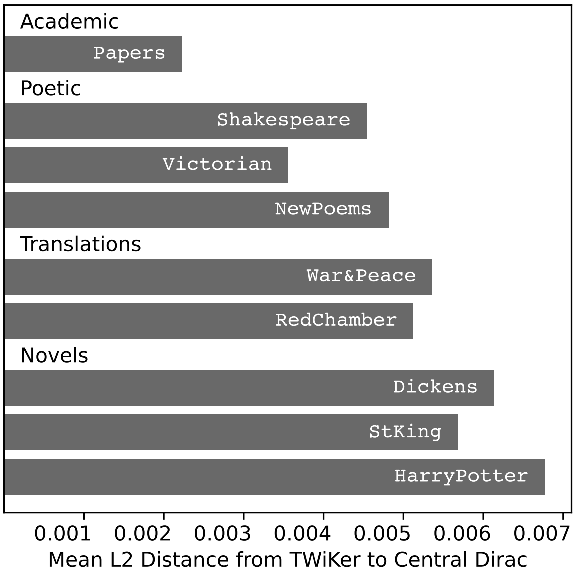


Figure 4: Mean deviation of learned `TWiKers` from the Central Dirac $[0, 1, 0]$ across nine corpora. Higher values suggest broader attention spread at the lexical level, often associated with more dynamic or loosely structured prose. Lower values indicate tighter, more self-contained word usage, reflecting semantically denser expression or a more formal tone.

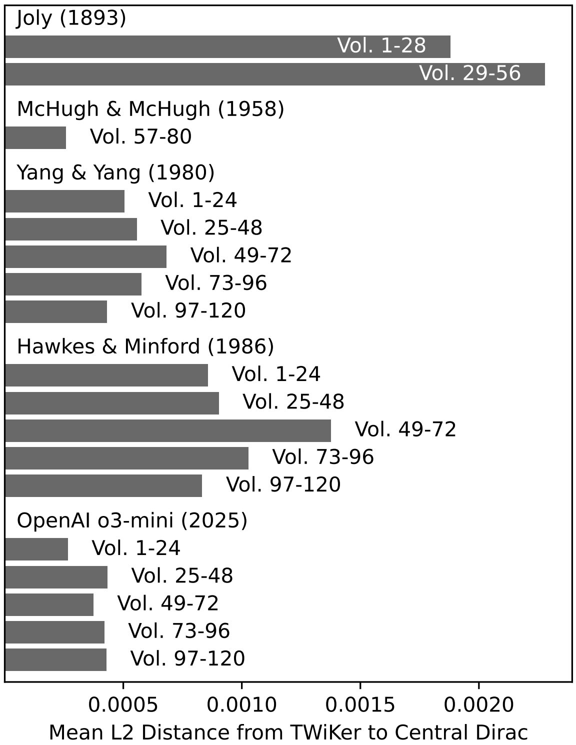


Figure 5: Mean deviation of learned `TWiKers` from Central Dirac $[0, 1, 0]$ across five English translations of *The Dream of the Red Chamber*.

age `TWiKer` deviation by PoS tags in each corpus and apply KMeans clustering. As shown in Figure 6a, clustering with all five translations is nearly perfect: the only exception is McHugh (M57), which groups with the AI translation, and the two Joly corpora (J1, J29) are separated to satisfy the five-cluster constraint. When we exclude the AI translation, Figure 6b shows that all corpora are clustered precisely. As two baselines, we also cluster based on PoS tag distributions (Figure 6c) and token embeddings averaged across PoS tags (Figure 6d). Both baselines lack sufficient granularity, resulting in considerable mixing among human-translated versions

4.4 Ablation Study

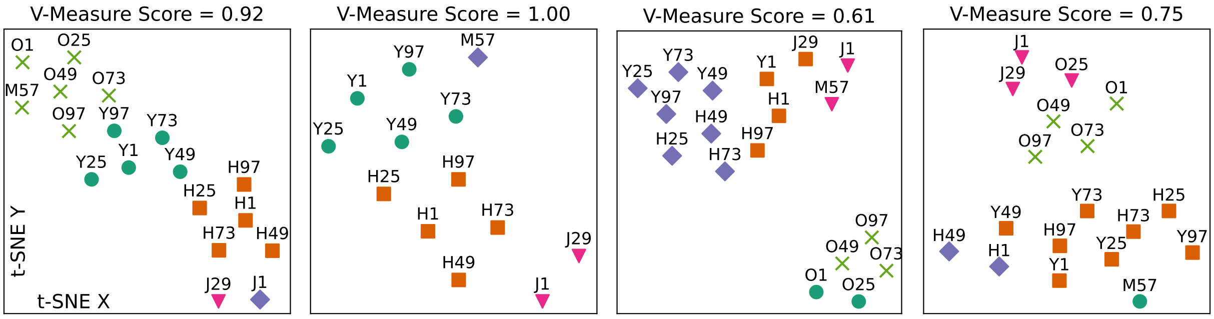
To assess the impact of architectural choices on `TWiKer` behavior, we conduct ablations on three factors: kernel size, whether `TWiKers` are applied to keys or values, and whether they vary across attention heads. We observe that overall lexical patterns–such as content words being self-focused and function words distributing attention–remain consistent across different configurations. Head-specific `TWiKers`, which scale with the number of attention heads, can smooth deviation patterns and further amplify improvements in training dynamics. The key findings are summarized here, with full details and additional plots available in Appendix D..

5 Training-focused Experiments

We have demonstrated that `TWiKers` effectively encode both lexical and stylistic characteristics of human language by modulating local attention spread. This interpretable mechanism indicates the potential of `TWiKers` to serve as a beneficial inductive bias for model training. In this section, we integrate `TWiKers` into LLaMA-3 (Touvron et al., 2023) and evaluate their impact on training dynamics.

We conduct experiments across all GLUE benchmark tasks. Due to computational constraints, we adopt LoRA (Hu et al., 2021) for finetuning LLaMA-3-8B, with a rank of 16. All tasks use a fixed learning rate of 10^{-4}, and models are trained for three epochs.

We consider three `TWiKer` configurations: (1) **OFF**, where `TWiKers` are disabled, and approximately 6.8 million parameters are updated via LoRA; (2) **SMALL**, where `TWiKers` with kernel size 3 are applied to values only and shared across heads, introducing roughly 0.4 million additional parameters (vocabulary size $\times$ 3); and (3) **LARGE**, where `TWiKers` with kernel size 5 are applied to



(a) TWiKer-based (b) TWiKer-based, no OpenAI (c) PoS-based (d) Embedding-based

Figure 6: Clustering five English translations of *The Dream of the Red Chamber*. Each point represents one corpus (∼24 chapters), where the label shows the ground-truth (initial of the first translator's name and the starting chapter number; see Figure 5), and the marker shape indicates clustering results. We use a simple KMeans algorithm, starting from 100 different random states, and show the best results as above. Subfigures (a) and (b) are based on learned TWiKer weights, and (c) and (d), as baselines, are respectively based on PoS tag distributions and token embeddings averaged across PoS tags.

Task	Metric	OFF	SMALL	LARGE
rte	Loss	0.5037	0.4901	**0.4704**
	Acc	0.8342	0.8339	**0.8628**
mrpc	Loss	0.3821	**0.3805**	0.3863
	Acc	0.8625	**0.8701**	0.8652
	F1	0.9054	**0.9062**	0.9037
stsb	Loss	0.4182	0.3999	**0.3862**
	PC	0.9042	0.9072	**0.9107**
	SC	0.9069	0.9072	**0.9125**
cola	Loss	**0.3804**	0.4132	0.3812
	MC	0.6830	0.6473	**0.6878**
sst2	Loss	0.1722	0.1873	**0.1580**
	Acc	0.9667	0.9633	**0.9690**
qnli	Loss	0.1913	0.1925	**0.1848**
	Acc	**0.9573**	0.9568	0.9553
qqp	Loss	0.2977	0.3029	**0.2959**
	Acc	0.9111	**0.9191**	0.9189
	F1	0.8913	**0.8919**	0.8916
mnli	Loss	0.3651	**0.3555**	0.3678
	Acc	0.9148	**0.9155**	0.9111

Table 2: GLUE Benchmark with LLaMA-3-8B and TWiKer. **MC** = Matthews Correlation, **PC** = Pearson Correlation, **SC** = Spearman Correlation. WNLI is excluded due to accuracy lower than random chance.

both keys and values with head-specific variation, introducing around 41 million parameters (vocabulary size × 5 × 2 × number of heads).

As shown in Table 2, introducing TWiKers generally improves task performance compared to the **OFF** baseline. In most cases, the **LARGE** configuration achieves the best results, though **SMALL** TWiKers can sometimes be competitive despite their minimal parameter overhead. These results suggest that TWiKers can offer an effective and interpretable inductive bias to enhance transformer-based models across diverse language tasks.

Nevertheless, our experiments leverage LoRA, whose parameter scale is comparable to that of TWiKers. As a result, the observed improvements may partly stem from interactions between LoRA and TWiKers. In full-parameter pretraining or finetuning scenarios, where the overhead introduced by TWiKers is negligible, their influence may be less pronounced or manifest differently, demanding further large-scale experiments.

6 Conclusion

We have introduced TWiKers, a novel mechanism that equips transformers with token-specific convolutional kernels, providing a lightweight inductive bias toward vicinity reliance—an inherent property of human languages. Our language-focused experiments show that TWiKers capture meaningful lexical and syntactic behaviors without supervision: content words retain self-focus, while function words redistribute attention to neighboring tokens. This generalizes across diverse English corpora, reflecting both low-level linguistic regularities and high-level stylistic variation. Such interpretability naturally suggests that TWiKers may benefit model training, as partly demonstrated by our downstream finetuning experiments. As the first directly interpretable transformer weights, TWiKers may inspire new directions for linguistic analysis and the development of efficient, interpretable neural weights for language modeling.

Limitations

Our study has two main limitations. First, tokens do not always correspond to words under modern subword tokenization schemes. We address this by excluding suffix tokens from our analysis and consistently aligning tokens to whole words, which reduces statistical power but preserves word-token alignment for most of the text. For more linguistically demanding tasks, pretraining with larger, word-based vocabularies may help. Second, due to resource constraints, our experiments use two small-scale LLMs: GPT-2 and LLaMA-3-8B (with LoRA). Although small, they retain the essential properties of causal decoder models and is suitable for testing our hypotheses. Also, our analysis is restricted to English. Extending `TWiKers` to languages with diverse morphological and syntactic features is an important direction for future work.

References

Enan Ajmain. 2022. Stephen king books dataset. `https://www.kaggle.com/datasets/lujar1762/stephen-king-books`. Kaggle dataset.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Re. 2024. Simple linear attention language models balance the recall-throughput tradeoff. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 1248–1261.

Sister Magdalen Louise Blum. 1950. The imagery in the poetry of gerard manley hopkins. Ma thesis, University of New Mexico.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics (ACL 2017)*, volume 5, pages 135–146.

Miriam Cha, Youngjune Gwon, and H. T. Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM 2017, pages 2003–2006, New York, NY, USA. Association for Computing Machinery.

Alison Chapman. 2022. Digital victorian periodical poetry project (dvpp). `https://dvpp.uvic.ca/`. University of Victoria, last accessed April 24, 2022.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. Longlora: Efficient fine-tuning of long-context large language models. In *The International Conference on Learning Representations (ICLR 2024)*.

George Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype bert embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Rodger Drew. 1996. *Symbolism and Sources in the Painting and Poetry of Dante Gabriel Rossetti*. Phd thesis, University of Glasgow.

Matthew S Dryer. 1992. The greenbergian word order correlations. *Language*, 68(1):81–138.

Wenyu Du, Zhouhan Lin, Yikang Shen, Timothy O'Donnell, Yoshua Bengio, and Yue Zhang. 2020. Exploiting syntactic structure for better language modeling: A syntactic distance approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6611–6628.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *Preprint*, arXiv:2407.21783.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM 2024)*.

Fritz Günther, Luisa Rinaldi, and Marco Marelli. 2019. Vector-space models for the representation of word meanings: a survey. *Language, Cognition and Neuroscience*, 34(5):572–590.

Anthony H. Harrison. 2004. Pre-raphaelite and ruskinian aesthetics. *The Victorian Web*.

Ali Hassani, Steven Walton, Jiacheng Li, Shengjia Li, and Humphrey Shi. 2022. Neighborhood attention transformer. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)*, pages 6185–6194.

David Hawkes and John Minford. 1986. *The Story of the Stone*. Penguin Books. Five-volume edition; Hawkes translated Chapters 1–80, Minford translated 81–120.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Ryan Holbrook. 2020. Scientific papers dataset. https://www.kaggle.com/datasets/ryanholbrook/scientific-papers. Kaggle dataset.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhong Xu, Brian Li, Lihong Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

John Dixon Hunt. 1968. *The Pre-Raphaelite Imagination, 1848-1900*. Routledge & Kegan Paul, London.

Internet Archive. 2020. The dream of the red chamber (book i). https://archive.org/details/the-dream-of-the-red-chamber-book-i_gutenberg-etext9603. Archive.org (Gutenberg text).

Jacob Jewusiak. 2021. Tennyson's wrinkled feet: Ageing and the poetics of decay. *19: Interdisciplinary Studies in the Long Nineteenth Century*, 2021(32):1–20.

H. Bencraft Joly. 1893. *The Dream of the Red Chamber*. Kelly & Walsh Press, Hong Kong; Macao Commercial Printing Bureau. English translation of Chapters 1–56.

Rupanshu Kapoor. 2024. Harry potter books dataset. https://www.kaggle.com/datasets/rupanshukapoor/harry-potter-books. Kaggle dataset.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 181–184. IEEE.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2018)*, pages 66–71.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211.

Tuan M. V. Le and Hady W. Lauw. 2017. Semantic visualization for short texts with word embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 2074–2080.

Patrick Lewis, Ethan Perez, Adam Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Kelvin Lu, Sebastian Riedel, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33.

Xuebo Li, Linyi Wu, Xinyang Li, and Cho-Jui Hsieh. 2022. Understanding transformers via n-gram statistics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 9950–9960.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling (COLM 2024)*.

Peng Liu, Weizhu Xu, Yu Zhang, Xingang Lin, Xuezhi Ma, Jie Liu, and Steven C.H. Hoi. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, pages 10012–10022.

P. N. Madhusudana. 2022. Dramatic monologues: A study of robert browning's narrative techniques. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 9(10):f47–f51.

Josh McAdams. 2020. Jane austen and charles dickens collection. https://www.kaggle.com/datasets/joshmcadams/jane-austin-and-charles-dickens. Kaggle dataset.

Florence McHugh and Isabel McHugh. 1958. *The Dream of the Red Chamber*. Pantheon Books, New York, NY. Translated from Franz Kuhn's German version.

Matt McKay. 2016. Text of war and peace. https://github.com/mmcky/nyu-econ-370/blob/master/notebooks/data/book-war-and-peace.txt. GitHub repository.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR 2013)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS 2013)*, volume 26.

Nicole Miras. 2024. Art, myth, and literature: The pre-raphaelites. *The Crossroads Gazette*.

Piero Molino, Yang Wang, and Jiawei Zhang. 2019. Parallax: Visualizing and understanding the semantics of embedding spaces via algebraic formulae. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 165–180, Florence, Italy. Association for Computational Linguistics (ACL 2019).

Aida Nematzadeh, Stephan C Meylan, and Thomas L Griffiths. 2017. Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, pages 859–864.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543.

Molly Petersen and Lonneke van der Plas. 2023. Can language models learn analogical reasoning? investigating training objectives and comparisons to human performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 16414–16425, Singapore. Association for Computational Linguistics.

Poetry Foundation. 2023. Poetryfoundation.org: Data summary. `https://www.poetryfoundation.org`. Public online archive.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, volume 32. Curran Associates, Inc.

Moussa Samir, Stephane Dufau, Gareth Gaskell, and Anastasia Ulicheva. 2020. Modeling semantic priming and lexical decision with distributed semantic spaces. In *Proceedings of the Cognitive Science Society*, volume 42, pages 1054–1060.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic (ACL 2016)*, pages 1715–1725.

Svetlana S. Takhtarova and Amelia Sh. Zubinova. 2018. The main characteristics of stephen king's idiostyle. *Vestnik Volgogradskogo gosudarstvennogo universiteta Serija 2 Jazykoznanije*, 17(3):139–147.

Jeuti Talukdar. 2024. Narrative techniques in the novels of charles dickens: A comparative analysis. *International Journal of Creative Research Thoughts (IJCRT)*, 12(2):154–160.

Binh Thai, Yu Wu, Pranav Jain, Ankur P Ravula, and Mohit Iyyer. 2020. N-grammer: Augmenting transformers with latent n-grams. *arXiv preprint arXiv:2007.12766*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, volume 30.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. *Proceedings of NAACL*, pages 1134–1145.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xianyi Yang and Gladys Yang. 1980. *A Dream of Red Mansions*. Foreign Languages Press, Beijing. Three-volume English translation of the full novel.

Howard Yen, Tianyu Gao, and Danqi Chen. 2024. Long-context language modeling with parallel context encoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2588–2610, Bangkok, Thailand. Association for Computational Linguistics.

Manzil Zaheer, Gokhan Guruganesh, Souvik Dubey, James Ainslie, Claudio Alberti, Santiago Ontanon, Paul Pham, Abhishek Ravula, Qifan Wang, and Li Yang. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 17283–17303.

Asim Zahid. 2021. Shakespeare plays dataset. https://www.kaggle.com/datasets/asimzahid/shakespeare-plays. Kaggle dataset.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. ClusterLLM: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920, Singapore. Association for Computational Linguistics (ACL 2023).

Xunjie Zhu, Tingfeng Li, and Gerard de Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637, Melbourne, Australia. Association for Computational Linguistics (ACL 2018).

A Data Declaration

We reviewed all nine corpora (see Table 1 for detailed descriptions of texts, time periods, and sources) to ensure that no personally identifying or offensive content was present. All materials are drawn from published literary works (public domain or widely distributed) and sampled scientific abstracts; we did not include private correspondence or unpublished personal data. Automated scripts scanned for full name patterns, email style strings, and offensive keywords; any hits were manually inspected and where necessary redacted. In the case of scientific articles, we also removed author bylines, institutional affiliations, and acknowledgments to protect anonymity.

All our data are in English. For originally non-English works (`War and Peace` and `The Dream of the Red Chamber`), we use their English translations; we also note multiple translator variants and demographic context (e.g. British vs. Russian vs. Chinese authors) in Table 1. For each corpus we record the number of works (e.g. 17 Shakespeare plays, 2 216 Victorian poems, 5 000 contemporary poems, 1 000 scientific article paragraphs, etc.), the source citation, and the predominant linguistic phenomena (e.g. inverted syntax and metaphor in Shakespeare, nominalization and passive constructions in scientific prose).

Across all corpora we processed approximately 1.2 million tokens. Each corpus was split at the document level into 80% train, and 20% test sets stratified by author and genre to preserve stylistic diversity. Detailed token counts per split (and per PoS tag) are provided in the supplementary Jupyter notebook, alongside document counts and PoS tag distributions.

B Engineering Details

We trained GPT-2 Base (117 M parameters) using a single NVIDIA V100 (40 GB) GPU. Total compute per corpus averaged under one GPU-hour (including both forward and backward passes), with all experiments running on the same V100 instance.

All main experiments reported in Section 4.1 used fixed hyperparameters: a learning rate of 5×10^{-5} for the Transformer weights and 5×10^{-3} for the `TWiKer` kernel parameters; a batch size of 6 for both training and evaluation; a `TWiKer` kernel size of 3 applied to the value projections in the attention mechanism; 2×30 training epochs; and a softmax temperature of 0.4 for normalizing `TWiKers`. Hyperparameter sweeps and ablation studies are discussed separately in Appendix D.

`TWiKers` are implemented through local modifications to Huggingface's `transformers` library. For data processing and analysis, we use SpaCy's `en_core_web_sm` model for part-of-speech tagging and NLTK's default rule-based tokenizer for sentence segmentation.

All results are reproducible via one-click experiment scripts and plotting utilities included in our released codebase and dataset package.

C TWiKers across PoS Tags in Corpora

In this Appendix, we present continued results for Section 4.2. Figure 7 shows the mean deviation of learned `TWiKers` from the Central Dirac kernel across nine corpora, broken down by PoS tags. These results reveal consistent trends in attention spread across lexical categories, while also highlighting stylistic variation among genres and time periods.

Charles Dickens and **Victorian Poetry** are the only corpora in which *prepositions* likely exhibit greater deviation than *determiners*. In Dickens, this may reflect a narrative style that tends to emphasize spatial density and rhythmic layering (Talukdar, 2024). For example, in *Great Expectations*:

> *"In a corner of the forge, the fire was burning brightly, and Joe was at his bellows, energetically puffing away."*

Here, *prepositions* such as "in", "of", and "at" likely function as structural anchors, distributing descriptive weight across the sentence. This style could be seen as aligning with Victorian literary aesthetics, where detailed spatial descriptions and atmospheric depth were common. `TWiKers` can

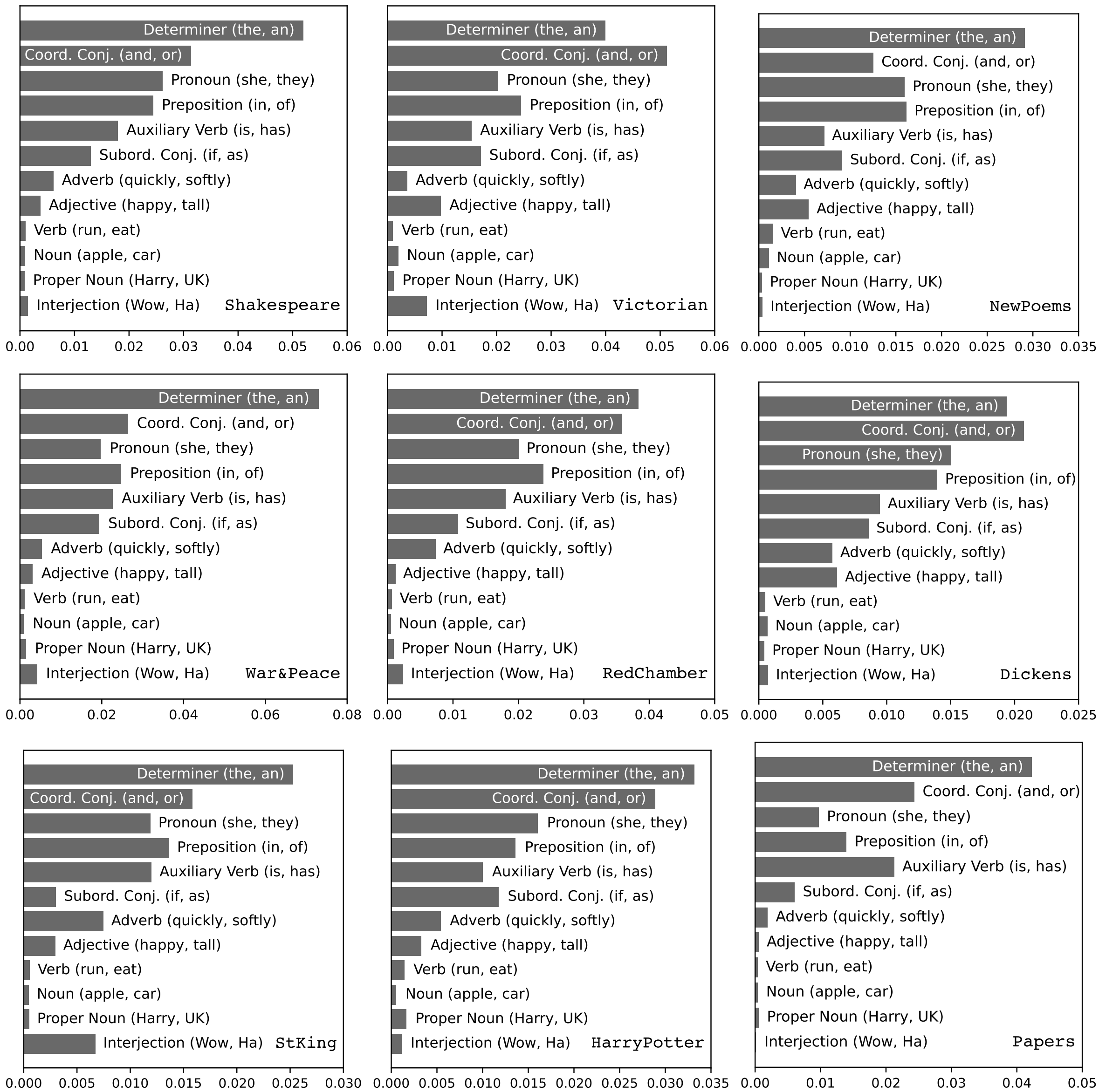


Figure 7: Mean deviation of learned TWiKers from Central Dirac $[0, 1, 0]$ across PoS tags in nine corpora.

learn to spread attention accordingly, capturing the rhetorical centrality of prepositional phrases in Dickens's prose.

Victorian poetry, though also showing elevated prepositional deviation, appears to follow a different stylistic rationale. Many literary scholars have noted that poets like Alfred Tennyson, Gerard Manley Hopkins, and Dante Gabriel Rossetti often favor determiner-noun imagery over clause-based narrative progression (Jewusiak, 2021; Blum, 1950; Drew, 1996). This stylistic choice likely reflects an emphasis on visual immediacy and symbolic precision, where *prepositions* often serve dual roles: indicating location and reinforcing prosodic balance. For instance, in Tennyson's *Tithonus*:

> *"The woods decay, the woods decay and fall..."*

Or Hopkins's *The Windhover*:

> *"The achieve of, the mastery of the thing!"*

Such usage suggests that *prepositions* and *determiners* function not merely as grammatical elements but as imagistic anchors. In contrast to narrative poets like Robert Browning, who rely heavily on *conjunctions* for logical progression ("And then she smiled..."), these poets emphasize stasis, vision, and repetition (Madhusudana, 2022). This static and visual emphasis connects closely with contemporary Victorian movements, such as the Pre-Raphaelite focus on symbolic and detailed visual imagery (Harrison, 2004; Hunt, 1968; Miras, 2024).

Additionally, **Victorian** poetry is the only corpus in which *determiners* likely deviate more than *conjunctions*. This could be attributed to the design of our dataset, which includes poets who often prioritize determiner-led imagery over logical connectives. For example, in Tennyson's *The Lady of Shalott*:

> *"The mirror crack'd from side to side;*
> *'The curse is come upon me,' cried*
> *The Lady of Shalott."*

Each instance of "the" may function as a visual or symbolic anchor–"mirror", "curse", "lady"–while *conjunctions* are comparatively minimized. This focus on determiner-led imagery is not universal among Victorian poets; for example, Browning and Christina Rossetti are known fo their reliance on clause-driven narrative progression. Our corpus likely foregrounds poets with a more determiner-centric style.

Stephen King presents a third, striking divergence: his is the only corpus where *interjections* appear to show the highest TWiKer deviation. This may be due to his focus on emotional immediacy, especially in horror and psychological suspense, where interjections often serve as narrative turning points (Takhtarova and Zubinova, 2018). From *The Green Mile*:

> *"We each owe a death, there are no exceptions, I know that, but sometimes, oh God, the Green Mile is so long."*

And in *Carrie*:

> *"No. Oh dear God, please no. (please let it be a happy ending)"*

These utterances do not carry strict syntactic function, but they likely help regulate pacing, convey fear, and anchor character perspective. TWiKers may capture this by assigning wider attention to such tokens, reflecting their dependence on surrounding discourse rather than immediate syntactic neighbors.

Taken together, these stylistically grounded deviations could support a key claim: TWiKers do not merely encode syntactic proximity–they can internalize genre conventions, authorial style, and literary tradition. The model's attention behavior highly resonates with deep patterns in English literary history, offering an interpretable bridge between data-driven learning and humanistic reading.

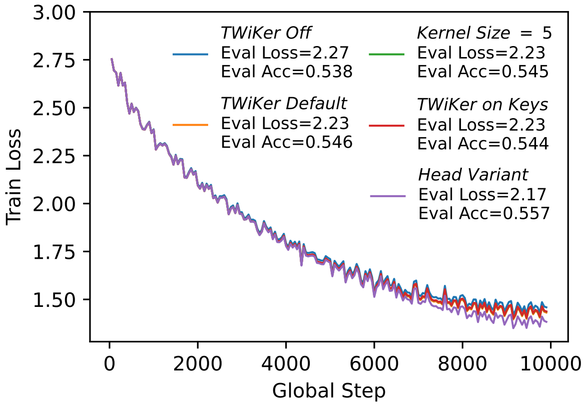


Figure 8: Training loss curves for ablation variants of TWiKer on the HarryPotter corpus. *TWiKer Default* corresponds to applying TWiKers with kernel size 3 to the values in the attention mechanism, shared across all attention heads. Final evaluation loss and accuracy are reported in the legend.

D Ablation Study

In this section, we examine how various architectural choices influence the behavior of TWiKers, using the HarryPotter corpus. The default configuration uses a kernel size of 3, with TWiKers applied to the values in the attention mechanism, shared across all attention heads. This setup underpins the results presented in Section 4.1 and Section 4.2.

We consider three ablation variants, each modifying a single factor while keeping all others fixed:

- **Kernel size = 5**: Increases the TWiKer kernel width, allowing tokens to incorporate a broader local context.
- **TWiKer on Keys**: Applies TWiKers to the keys instead of the values, shifting the locality bias from the value aggregation to the query-side matching process.
- **Head Variant**: Assigns a separate TWiKer to each attention head within the input layer, enabling head-specific attention patterns.

As a baseline, we also train the base model with TWiKer deactivated.

Figure 8 shows the training loss curves for each configuration. The introduction of TWiKers adds only a small number of parameters, resulting in negligible disruption to the optimization process, while offering slight improvements in loss and accuracy. However, when allowed to vary by head (*Head Variant*), we observe slight improvements in both convergence rate and final evaluation accuracy. This suggests that `TWiKers` can serve as a lightweight and semantically grounded inductive bias in language modeling. Nevertheless, as noted in Limitations, all results are based on GPT-2. We do not claim general efficiency or scalability of `TWiKers` at larger model scales, and leave this for future investigation.

Figure 9 shows the mean deviation of learned `TWiKer` kernels from the Central Dirac across PoS tags under different configurations. Across all these variants, the overall pattern holds: function words (e.g., determiners, conjunctions) tend to shift attention to neighbors, while content words (e.g., nouns, verbs) retain self-focus. Increasing the kernel size to five leads to broader deviation, especially for function words. Subordinate conjunctions show an outstanding relative increase in deviation when `TWiKers` are applied to keys, likely because their clause-linking function interacts more strongly with the query-side of attention. Allowing variation across heads (*Head Variant*) results in smoother distance distributions across PoS categories, suggesting a regularizing effect from distributing the locality pattern across multiple attention paths.

E Use of AI Assistants

We used ChatGPT-4o and DeepSeek R1 to help write Python code and improve sentences. No part of the code or paper was generated by AI without human guidance and verification.

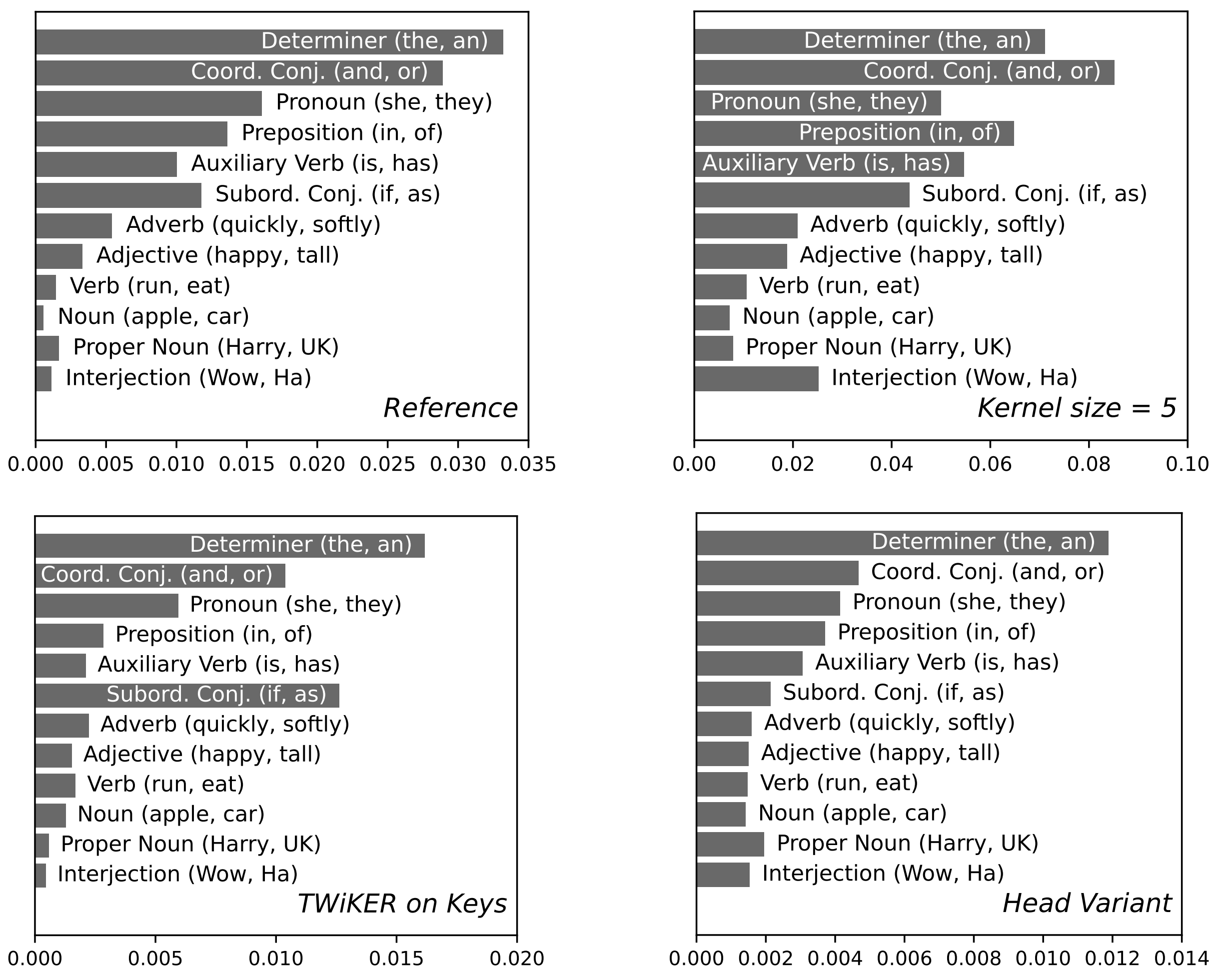


Figure 9: Mean deviation of learned `TWiKers` from the Central Dirac $[0, 1, 0]$ across PoS tags for different architectural configurations. *Reference*: kernel size = 3, `TWiKer` on values, head-invariant.