Novelty Assessment Report

Paper: Decoupling Positional and Symbolic Attention in Transformers

PDF URL: https://openreview.net/pdf?id=V38yAoqddQ

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-04

Abstract

An important aspect subtending language understanding and production is the ability to independently encode positional and symbolic information of the words within a sentence. In Transformers, positional information is typically encoded using Positional Encodings (PEs). One such popular PE, namely Rotary PE (RoPE), has been widely used due to its empirical success. Recently, it has been argued that part of RoPE's success emerges from its ability to encode robust positional and semantic information using large and small frequencies, respectively. In this work, we perform a deeper dive into the positional versus symbolic dichotomy of attention heads behavior, both at the theoretical and empirical level. We provide general definitions of what it means for a head to behave positionally or symbolically, prove that these are two mutually exclusive behaviors and develop a metric to quantify them.

We apply our framework to analyze Transformer-based LLMs using RoPE and find that all heads exhibit a strong correspondence between behavior and frequency use.

Finally, we introduce canonical tasks designed to be either purely positional or symbolic, and demonstrate that the Transformer performance causally relates to the ability of attention heads to leverage the appropriate frequencies. In particular, we show that we can control the Transformer performance by controlling which frequencies the attention heads can access. Altogether, our work provides a detailed understanding of RoPE, and how its properties relate to model behavior.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Decoupling Positional and Symbolic Attention Mechanisms in Transformers**

A total of **15 papers** were analyzed and organized into a taxonomy with **11 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Positional Encoding Design and Analysis**
- **Disentangled Attention Mechanisms**
- **Application-Specific Transformer Architectures**
- **Specialized Domain Applications**

Complete Taxonomy Tree

- Decoupling Positional and Symbolic Attention Mechanisms in Transformers Survey Taxonomy
- Positional Encoding Design and Analysis
    - Rotary Positional Encoding (RoPE) Analysis ★ (2 papers)
    - [0] Decoupling Positional and Symbolic Attention in Transformers (Anon et al., 2026) View paper
    - [1] Decoupling Positional and Symbolic Attention Behavior in Transformers (Felipe Urrutia, 2025) View paper
    - Alternative Positional Encoding Methods (2 papers)
    - [2] Relative-position embedding based spatially and temporally decoupled Transformer for action recognition (Yujun Ma, 2024) View paper
    - [11] A Simple and Effective Positional Encoding for Transformers (Bhojanapalli, 2021) View paper
    - Domain-Specific Positional Encoding Adaptations (2 papers)
    - [10] HRPE: Hierarchical Relative Positional Encoding forÂ Transformer-Based Structured Symbolic Music Generation (Pengfei Li, 2024) View paper
    - [14] Span-adaptive Transformer for the Cascade Relation Triple Extraction (Kai Liu, 2021) View paper
- Disentangled Attention Mechanisms
    - Content-Position Disentanglement in Language Models (1 papers)
    - [4] Deberta: Decoding-enhanced bert with disentangled attention (Pengcheng He, 2020) View paper
    - Multimodal Feature Decoupling (2 papers)
    - [9] Multivariate Diffusion Transformer with Decoupled Attention for High-Fidelity Mask-Text Collaborative Facial Generation (Yushe Cao, 2025) View paper
    - [13] A Transformer-Based Decoupled Attention Network for Text Recognition in Shopping Receipt Images (Lang Ren, 2021) View paper
- Application-Specific Transformer Architectures
    - Structured Symbolic Recognition (1 papers)
    - [6] PosFormer: Recognizing Complex Handwritten Mathematical Expression with Position Forest Transformer (Guan, 2024) View paper
    - Vision and Spectral-Spatial Classification (1 papers)
    - [3] End-to-end convolutional network and spectral-spatial Transformer architecture for hyperspectral image classification (Shiping Li, 2024) View paper
    - Sequence Generation with Structural Constraints (2 papers)

    - [7] A Transformer-based Function Symbol Name Inference Model from an Assembly Language for Binary Reversing (Kim Hyun-Jin, 2023) View paper
    - [8] MixSong: Diverse and Strictly Formatted Chinese Poetry Generation (Xinglong Song, 2025) View paper
- Specialized Domain Applications
    - Molecular and Chemical Property Prediction (1 papers)
    - [12] Deep peak property learning for efficient chiral molecules ECD spectra prediction (Li Hao, 2024) View paper
    - Financial Risk and Cross-Sector Analysis (1 papers)
    - [5] An Empirical Analysis of the Impact of ESG Management Strategies on the Long-Term Financial Performance of Listed Companies in the Context of China â¦ (D Liu, 2025) View paper
    - Logical and Relational Architectures (1 papers)
    - [15] A logical re-conception of neural networks: Hamiltonian bitwise part-whole architecture (Granger, n.d.) View paper

Narrative

Core task: Decoupling positional and symbolic attention mechanisms in Transformers. The field centers on understanding and improving how Transformers encode position information separately from content-based (symbolic) attention. The taxonomy reveals four main branches: Positional Encoding Design and Analysis examines foundational schemes such as rotary positional encoding (RoPE) and relative position methods, exploring their mathematical properties and limitations; Disentangled Attention Mechanisms investigates architectures that explicitly separate positional and content-based computations, as seen in works like DeBERTa[4] and related decoupled designs; Application-Specific Transformer Architectures adapts these principles to particular tasks such as vision, time series, or structured prediction; and Specialized Domain Applications extends the ideas to niche settings including music generation, multivariate forecasting, and document understanding. Representative studies like PosFormer[6] and HRPE[10] illustrate how positional encoding choices directly shape model expressiveness, while others such as Relative Position Spatiotemporal[2] and Convolutional Spectral Spatial[3] blend positional reasoning with domain-specific inductive biases.

A particularly active line of work focuses on analyzing and refining RoPE-based encodings, where researchers probe how rotary embeddings interact with attention scores and whether they can be further disentangled to improve interpretability or generalization. Decoupling Positional Symbolic[0] sits squarely within this RoPE analysis cluster, closely aligned with Decoupling Positional Symbolic Behavior[1], which also examines the interplay between positional and symbolic components in rotary schemes. Compared to broader disentangled attention studies like DeBERTa[4] or Decoupled Attention Receipt[13], which propose architectural changes across multiple layers, the original paper emphasizes a more focused investigation of how RoPE's geometric structure can be decomposed and understood. This contrasts with application-driven works such as MixSong[8] or Multivariate Diffusion Decoupled[9], which prioritize task-specific performance over mechanistic insights. Overall, the work contributes to a growing effort to make positional encoding more transparent and controllable within the Transformer framework.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Decoupling Positional and Symbolic Attention Behavior in Transformers

Authors: Felipe Urrutia, Jorge Salas, Alexander Kozachinskiy, Cristian Buc Calderon, Hector Pasten, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

An important aspect subtending language understanding and production is the ability to independently encode positional and symbolic information of the words within a sentence. In Transformers, positional information is typically encoded using Positional Encodings (PEs). One such popular PE, namely Rotary PE (RoPE), has been widely used due to its empirical success. Recently, it has been argued that part of RoPE's success emerges from its ability to encode robust positional and semantic informati...

⚠ Similarity Notice

This paper is highly similar to the original paper; it may be a variant or near-duplicate. Please manually verify.

Contributions Analysis

Overall novelty summary. The paper provides formal definitions distinguishing positional from symbolic attention head behavior in Transformers using RoPE, alongside a novel metric to quantify this dichotomy. It resides in the 'Rotary Positional Encoding (RoPE) Analysis' leaf, which contains only two papers including this one. This represents a relatively sparse research direction within the broader taxonomy of 15 papers across multiple branches. The focused scope suggests the work addresses a specific gap in understanding RoPE's internal mechanisms rather than competing in a crowded subfield.

The taxonomy reveals that RoPE analysis sits within 'Positional Encoding Design and Analysis', adjacent to 'Alternative Positional Encoding Methods' (2 papers) and 'Domain-Specific Positional Encoding Adaptations' (2 papers). Neighboring branches include 'Disentangled Attention Mechanisms' (3 papers) and various application-specific architectures. While disentangled attention work like DeBERTa explicitly separates content and position through architectural changes, this paper takes a mechanistic approach to understanding how RoPE implicitly achieves separation through frequency allocation. The taxonomy structure indicates the field is exploring both architectural innovations and analytical frameworks in parallel.

Among 22 candidates examined across three contributions, none were found to clearly refute the paper's claims. The formal definitions of positional versus symbolic behavior examined 2 candidates with no refutations. The novel metric contribution examined 10 candidates, again with no overlapping prior work identified. The canonical task design similarly examined 10 candidates without finding substantial precedent. These statistics suggest that within the limited search scope, the paper's specific combination of formal analysis, quantification metrics, and controlled experiments appears relatively unexplored, though the search scale leaves open the possibility of relevant work beyond the top-22 semantic matches.

Based on the limited literature search of 22 candidates, the work appears to occupy a distinct analytical niche within RoPE research. The sparse taxonomy leaf and absence of refuting candidates suggest novelty in the specific mechanistic framework proposed. However, the modest search scope means this assessment reflects top-K semantic similarity rather than exhaustive field coverage, and related theoretical work on attention mechanisms may exist outside the examined set.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Formal definitions of positional and symbolic attention head behavior

Description: The authors introduce mathematical definitions characterizing when an attention head acts positionally (logits invariant under key vector permutations) versus symbolically (logits equivariant under key vector permutations). They prove these behaviors are mutually exclusive unless attention is uniform, and show certain operations require one behavior but not the other.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Attention as Binding: A Vector-Symbolic Perspective on Transformer Reasoning

URL: View paper

Brief Assessment

Attention as Binding[26] focuses on interpreting attention through vector symbolic architectures (VSAs) as binding/unbinding operations, not on formal definitions distinguishing positional versus symbolic attention head behavior based on logit invariance/equivariance under key permutations.

2. Decoupling Positional and Symbolic Attention Behavior in Transformers

URL: View paper

Brief Assessment

Decoupling Positional Symbolic Behavior[1] appears to be the same work as the original paper (identical authors, content, and structure), not prior work that could refute novelty.

Contribution 2: Novel metric quantifying positional and symbolic behavior of attention heads

Description: The authors develop a metric that assigns positional and symbolic scores to attention heads at various granularities, from specific inputs and frequencies to per-head characterization. This enables visualization of model behavior in a positional-symbolic plane and reveals sharp correspondence between RoPE frequencies and head behavior types.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Unveiling visual perception in language models: An attention head analysis approach

URL: View paper

Brief Assessment

Visual Perception Attention[17] focuses on analyzing attention heads in multimodal models for visual token processing, not on quantifying positional versus symbolic behavior in language-only transformers with RoPE.

2. On the token distance modeling ability of higher RoPE attention dimension

URL: View paper

Brief Assessment

Token Distance RoPE[22] focuses on identifying 'positional heads' that capture long-range dependencies through dimension-level correlation analysis, rather than developing a metric to quantify both positional and symbolic behaviors across different granularities as in the original work.

3. Attention speaks volumes: Localizing and mitigating bias in language models

URL: View paper

Brief Assessment

Attention Localizing Bias[18] focuses on quantifying entity preference and bias localization in LLMs through attention analysis, not on characterizing positional versus symbolic attention behavior at different granularities or RoPE frequency correspondence.

4. How does attention work in vision transformers? A visual analytics attempt

URL: View paper

Brief Assessment

Attention Vision Transformers[25] focuses on vision transformers (ViTs) and introduces pruning-based metrics for head importance and autoencoder-based learning for attention patterns in image patches. It does not address positional versus symbolic behavior dichotomy in language models or develop metrics for quantifying these behaviors across RoPE frequencies.

5. Stochastic subnetwork induction for contextual perturbation analysis in large language model architectures

URL: View paper

Brief Assessment

Stochastic Subnetwork Induction[19] focuses on contextual perturbation analysis through stochastic subnetwork induction, not on developing metrics to quantify positional versus symbolic attention head behavior or analyzing RoPE frequency correspondence.

6. Silent grammars in emergent language models: An exploratory study of latent instructional drift via stochastic scaffold morphogenesis

URL: View paper

Brief Assessment

Silent Grammars Emergent[20] focuses on scaffold recurrence and alignment stability indices in emergent language models, not on metrics for quantifying positional versus symbolic attention head behavior in transformers with RoPE.

7. Efficient Prompt Compression with Evaluator Heads for Long-Context Transformer Inference

URL: View paper

Brief Assessment

Evaluator Heads Compression[23] focuses on identifying attention heads that select important tokens for prompt compression in long-context inference, not on quantifying positional versus symbolic behavior patterns or analyzing RoPE frequency relationships.

8. Going where, by whom, and at what time: Next location prediction considering user preference and temporal regularity

URL: View paper

Brief Assessment

Next Location Prediction[24] focuses on human mobility prediction using multi-head attention for arrival time estimation in location forecasting, not on analyzing or quantifying positional versus symbolic behavior of attention mechanisms in transformers.

9. Direct visual grounding by directing attention of visual tokens

URL: View paper

Brief Assessment

Direct Visual Grounding[16] focuses on supervising attention of language tokens to visual tokens in vision-language models using KL divergence loss. This is fundamentally different from quantifying positional versus symbolic behavior of attention heads in language-only transformers with RoPE.

10. Unveiling simplicities of attention: Adaptive long-context head identification

URL: View paper

Brief Assessment

Adaptive Long Context[21] focuses on identifying local vs. long-context heads using second-order statistics, not on quantifying positional vs. symbolic behavior or analyzing RoPE frequency usage patterns.

Contribution 3: Canonical tasks demonstrating causal relationship between frequency access and performance

Description: The authors design intrinsically positional (Index task) and symbolic (Information Retrieval task) tasks, proving theoretically that pure positional heads cannot solve symbolic tasks and vice versa. They show experimentally that controlling which RoPE frequencies heads can access directly controls model performance, with characteristic U-shaped and inverted-U-shaped accuracy patterns emerging.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Frequency Effects on Syntactic Rule Learning in Transformers

URL: View paper

Brief Assessment

Frequency Effects Syntactic[33] studies frequency effects on syntactic rule learning in transformers, focusing on subject-verb agreement and word frequency in training data. This differs from the original paper's focus on RoPE frequency bands and their causal relationship to positional versus symbolic task performance.

2. Cuff-less Arterial Blood Pressure Waveform Synthesis from Single-site PPG using Transformer & Frequency-domain Learning

URL: View paper

Brief Assessment

Cuffless Blood Pressure[34] focuses on blood pressure waveform synthesis using transformers for medical signal processing, not on analyzing the relationship between RoPE frequencies and transformer performance on positional versus symbolic tasks.

3. HL-ESViT: High-Low Frequency Efficient Spiking Vision Transformer

URL: View paper

Brief Assessment

HL-ESViT[35] focuses on spiking neural networks and vision transformers with high-low frequency pathways for image processing, not on analyzing causal relationships between RoPE frequency access and transformer performance on positional versus symbolic tasks.

4. Harmonic Frequency-Separable Transformer for Instrument-Agnostic Music Transcription

URL: View paper

Brief Assessment

Harmonic Frequency Separable[30] focuses on music transcription using frequency-separable transformers for harmonic structure, not on causal relationships between RoPE frequency access and transformer performance on positional versus symbolic tasks.

5. HRFT: Mining High-Frequency Risk Factor Collections End-to-End via Transformer

URL: View paper

Brief Assessment

HRFT[32] focuses on mining formulaic risk factors in quantitative trading using transformers for symbolic regression, not on analyzing the relationship between RoPE frequency access and transformer performance on positional versus symbolic tasks.

6. Decoding stress specific transcriptional regulation by causality aware Graph-Transformer deep learning

URL: View paper

Brief Assessment

Causality Aware Graph[31] focuses on transcriptional regulation in biological stress responses using graph-transformer architectures, not on transformer attention mechanisms, RoPE frequencies, or positional versus symbolic task performance.

7. KV-Latent: Dimensional-level KV Cache Reduction with Frequency-aware Rotary Positional Embedding

URL: View paper

Brief Assessment

KV-Latent[27] focuses on KV cache compression through dimensional reduction and modifying RoPE frequency sampling for stability, not on designing canonical tasks to prove causal relationships between frequency access and transformer performance on positional versus symbolic tasks.

8. Causality-aware transformer networks for robotic navigation

URL: View paper

Brief Assessment

Causality Aware Navigation[28] focuses on causal relationships in robotic navigation environments (state transitions and action selection), not on the relationship between RoPE frequency access and transformer performance on positional versus symbolic tasks.

9. Decoupling Positional and Symbolic Attention Behavior in Transformers

URL: View paper

Brief Assessment

Decoupling Positional Symbolic Behavior[1] is the same paper as the original, containing identical task definitions and experimental results, thus cannot serve as prior work to refute the original's novelty claims.

10. U-shaped transformer with frequency-band aware attention for speech enhancement

URL: View paper

Brief Assessment

Frequency Band Attention[29] focuses on speech enhancement using frequency-band aware attention in transformers for audio processing, not on analyzing causal relationships between RoPE frequency access and transformer performance on positional versus symbolic tasks.

Appendix: Text Similarity Detection

Textual similarity detection checked 21 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Decoupling Positional and Symbolic Attention Behavior in Transformers

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Decoupling Positional and Symbolic Attention in Transformers View paper
- [1] Decoupling Positional and Symbolic Attention Behavior in Transformers View paper
- [2] Relative-position embedding based spatially and temporally decoupled Transformer for action recognition View paper
- [3] End-to-end convolutional network and spectral-spatial Transformer architecture for hyperspectral image classification View paper
- [4] Deberta: Decoding-enhanced bert with disentangled attention View paper
- [5] An Empirical Analysis of the Impact of ESG Management Strategies on the Long-Term Financial Performance of Listed Companies in the Context of China â¦ View paper
- [6] PosFormer: Recognizing Complex Handwritten Mathematical Expression with Position Forest Transformer View paper
- [7] A Transformer-based Function Symbol Name Inference Model from an Assembly Language for Binary Reversing View paper
- [8] MixSong: Diverse and Strictly Formatted Chinese Poetry Generation View paper
- [9] Multivariate Diffusion Transformer with Decoupled Attention for High-Fidelity Mask-Text Collaborative Facial Generation View paper
- [10] HRPE: Hierarchical Relative Positional Encoding forÂ Transformer-Based Structured Symbolic Music Generation View paper
- [11] A Simple and Effective Positional Encoding for Transformers View paper
- [12] Deep peak property learning for efficient chiral molecules ECD spectra prediction View paper
- [13] A Transformer-Based Decoupled Attention Network for Text Recognition in Shopping Receipt Images View paper
- [14] Span-adaptive Transformer for the Cascade Relation Triple Extraction View paper
- [15] A logical re-conception of neural networks: Hamiltonian bitwise part-whole architecture View paper
- [16] Direct visual grounding by directing attention of visual tokens View paper
- [17] Unveiling visual perception in language models: An attention head analysis approach View paper
- [18] Attention speaks volumes: Localizing and mitigating bias in language models View paper
- [19] Stochastic subnetwork induction for contextual perturbation analysis in large language model architectures View paper
- [20] Silent grammars in emergent language models: An exploratory study of latent instructional drift via stochastic scaffold morphogenesis View paper
- [21] Unveiling simplicities of attention: Adaptive long-context head identification View paper
- [22] On the token distance modeling ability of higher RoPE attention dimension View paper
- [23] Efficient Prompt Compression with Evaluator Heads for Long-Context Transformer Inference View paper
- [24] Going where, by whom, and at what time: Next location prediction considering user preference and temporal regularity View paper
- [25] How does attention work in vision transformers? A visual analytics attempt View paper
- [26] Attention as Binding: A Vector-Symbolic Perspective on Transformer Reasoning View paper
- [27] KV-Latent: Dimensional-level KV Cache Reduction with Frequency-aware Rotary Positional Embedding View paper
- [28] Causality-aware transformer networks for robotic navigation View paper
- [29] U-shaped transformer with frequency-band aware attention for speech enhancement View paper
- [30] Harmonic Frequency-Separable Transformer for Instrument-Agnostic Music Transcription View paper
- [31] Decoding stress specific transcriptional regulation by causality aware Graph-Transformer deep learning View paper
- [32] HRFT: Mining High-Frequency Risk Factor Collections End-to-End via Transformer View paper
- [33] Frequency Effects on Syntactic Rule Learning in Transformers View paper
- [34] Cuff-less Arterial Blood Pressure Waveform Synthesis from Single-site PPG using Transformer & Frequency-domain Learning View paper
- [35] HL-ESViT: High-Low Frequency Efficient Spiking Vision Transformer View paper