PhaseFormer: A Continuous-Phase Language Model Without Real-Valued Matrix Multiplication

Wu Yong, Wu Yuheng

2026-08-21

ChinaRxiv

Abstract

Current language models rely on real-valued matrix multiplication and gradient descent, consuming substantial computational resources. This paper proposes PhaseFormer, a language model architecture whose core computation is based on phase-difference operations. Its core computations are: (1) replacing dot products with phase-difference cosine sums, ∑cos(qi−ki); (2) using Adam to update angles in the continuous phase space (0~2π), rather than real-valued weights; and (3) replacing real-valued linear combinations with circular means, atan2(∑wi sin vi, ∑wi cos vi), for information aggregation. On V100, a 2-layer 128-dimensional PhaseFormer achieves: 0.435 in MLM cloze completion (rank 0.140 / top10 0.690), 0.430 in causal language modeling (rank 0.183 / top10 0.630), 0.315 in specialized reverse-causal training (rank 0.174 / top10 0.630), and 0.405 in reverse-causal zero-shot generalization (rank 0.156 / top10 0.635). Compared with a standard Transformer of the same scale (0.515), PhaseFormer reaches 84.5% of its performance. Probe experiments reveal a geometric identity in phase space: rotating the output phase by π produces a distance-complementarity relation, d(ϕ, wp[X]+π)=π−d(ϕ, wp[X]). Words from 6 semantic categories exhibit clustered distributions in phase space (intra-class distance < random distance, ratio 0.78~0.92), demonstrating that phase vectors encode semantic organizational information. This paper demonstrates the feasibility of a language modeling paradigm that replaces dot products with phase differences, providing a new path toward language models that move beyond real-valued matrix multiplication.

Full Text

PhaseFormer: A Continuous-Phase Language Model Without Real-Valued Matrix Multiplication

Wu Yong[1], Wu Yuheng[2]

[1] Shenzhen Shuangzi Tianxia Investment Co., Ltd., Shenzhen, China

[2] Tianjin University, Tianjin, China

August 21, 2026

Abstract

Current language models rely on real-valued matrix multiplication and gradient descent, consuming large amounts of computational resources. This paper proposes PhaseFormer, a language-model architecture whose core computation is based on phase-difference operations. Its core computations are: (1) replacing dot products with phase-difference residual chords and $\sum \cos(q_i - k_i)$; (2) using

Adam to update angles in the continuous phase space (0 2) rather than real-valued weights; and (3) replacing real-valued linear combinations with circular means atan2($\sum w_i \sin v_i, \sum w_i \cos v_i$) for information aggregation.

On a V100, a 2-layer, 128-dimensional PhaseFormer achieves: MLM cloze completion 0.435 (rank 0.140 / top10 0.690), causal language modeling 0.430 (rank 0.183 / top10 0.630), reverse-causal specialized training 0.315 (rank 0.174 / top10 0.630), and reverse-causal zero-shot generalization 0.405 (rank 0.156 / top10 0.635). Compared with a standard Transformer of the same scale (0.515), PhaseFormer reaches 84.5% of its performance.

Probing experiments reveal a geometric invariance in phase space: after rotating the output phase by π, a distance-complementarity relation is produced, $d(\phi, w_p[X] + \pi) = \pi - d(\phi, w_p[X])$. Words from six semantic categories show clustered distributions in phase space (within-class distance $<$ random distance, ratio 0.78–0.92), demonstrating that phase-vector encodings contain semantic organizational information.

This paper demonstrates the feasibility of a language-modeling paradigm that replaces dot products with phase differences, providing a new path toward language models that move beyond real-valued matrix multiplication.

Keywords: phase-difference attention, continuous phase space, matrix-multiplication-free, language model, wave interference

1 Introduction

Current mainstream language models (Transformer [1], GPT [3], LLaMA) rely on two core assumptions: (1) information is represented by real-valued vectors; and (2) learning updates these weights through gradient descent. These two assumptions lead to enormous computational overhead—matrix multiplication dominates the energy consumption of training and inference.

Complex numbers naturally contain two dimensions: amplitude (strength) and phase (direction). Real-valued systems use only the amplitude dimension, while the phase dimension has always been ignored. Existing works (iFairy [4], Fairy2i [5], qFHRR [6]) have explored using discrete phases as a means of compression, but they use phases to "approximate" real-valued operations, and in the aggregation step they return to real-valued linear combinations.

In contrast, this paper constructs a language model that does not rely on real-valued matrix multiplication in any component—from the underlying operators to the learning mechanism, all are based on phase-difference operations.

2 Model Architecture

2.1 Three Core Replacements

Table 1: Three Core Replacements: PhaseFormer vs. Standard Transformer

Component	Standard Transformer	PhaseFormer
Gradient update	Real-valued weights	Phase angles + modulo 2π constraint
Relation computation	Dot product $Q \cdot K^T/\sqrt{d}$	$\sum \cos(q_i - k_i)$
Information aggregation	Real-valued linear combination softmax $\cdot V$	Circular mean $\text{atan2}(\sum \sin, \sum \cos)$

Key distinction: iFairy/qFHRR made the first two replacements, but the third step returned to real-valued linear combinations. PhaseFormer replaces all three steps.

2.2 Attention Mechanism

Given phase queries Q and keys K (both floating-point angles of shape $[B, T, D]$), the correlation is computed as:

$$\text{sim} = \sum_i \cos(q_i - k_i) \qquad (1)$$

This is functionally the same as a dot product (measuring vector similarity), but without matrix multiplication—only phase differences and cosine operations.

Aggregation (weighted averaging in the angular domain, replacing the real-valued linear combination softmax $\cdot V$):

$$\text{out\_angle} = \text{atan2}\left(\sum_i w_i \sin(v_i), \sum_i w_i \cos(v_i)\right) \qquad (2)$$

where $w_i = \text{softmax}(\text{sim}/\sqrt{h})$, and h is the attention-head dimension (in implementation, $h = d/\text{num\_heads}$).

2.3 Positional Encoding and Model Configuration

The positional encoding is also a phase angle, added directly to the input phase:

$$x = \text{angle\_mod}(\text{token\_phase} + \text{pos\_phase}) \qquad (3)$$

Position and content are superposed in the same space (phase space), requiring no additional vector addition.

The word-phase prototype is `word_phases [80000, 128]`, with learnable continuous angles, 2 layers, dimension 128, 4 attention heads, and 10.5M parameters (a standard Transformer of the same scale has 10.7M).

3 Geometric Structure in Phase Space

3.1 Distance Complementarity under π Rotation

Probe experiments show that adding $+\pi$ to the output wave produces a complementary distance relationship:

$$d(\phi, w_p[X] + \pi) = \pi - d(\phi, w_p[X]) \qquad (4)$$

where $d(\cdot, \cdot)$ is the angular distance on the phase circle. This is a geometric identity in phase space: a phase shift of π sends the output to the antipodal point on the circle, maximizing its distance from the original point.

Important note: The flipped vector $w_p[X]+\pi$ does not necessarily correspond to semantic negation of some word in the vocabulary. The vocabulary embedding w_p is learned from random initialization and does not encode antonym-pair structure. We verified this point: among 200 sampled words, no word pair had a phase distance close to π (the proportion with $\left|d(\phi, w_p[X] + \pi) - \pi\right| < 0.5$ was 0%). Therefore, the π-flip operation should be understood as a distance-complement operator, not as a semantic-negation generator.

3.2 Input-Side π Rotation Is Absorbed (Rotation Equivariance)

Adding $+\pi$ to the entire input wave does not change the attention mechanism:

$$\cos((q + \pi) - (k + \pi)) = \cos(q - k) \qquad (5)$$

The interference consistency is 0.987 (0.998 when local $+\pi$ is at the `[MASK]` position). The input-side π rotation is absorbed by the attention mechanism and cannot serve as a semantic signal. This is consistent with the rotation equivariance of attention based on phase differences.

3.3 Semantic Clustering

The phase vectors of six semantic categories (animals/fruits/numbers/body/colors/transportation) exhibit clustered distributions in projection onto the complex plane:

For all six categories, the within-class distance is less than the random distance (ratio < 0.95), and all within-class distances are less than the between-class distances, demonstrating that phase-vector encoding captures semantic organizational information.

Figure 1: t-SNE projection of lexical waves for six semantic categories (red = animals, blue = fruits, green = numbers, orange = body, purple = colors, brown = transportation, gray = random words)

Figure 1: t-SNE projection of lexical waves for six semantic categories (red = animals, blue = fruits, green = numbers, orange = body, purple = colors, brown = transportation, gray = random words)

Figure 1: Figure 1: t-SNE projection of lexical waves for six semantic categories (red = animals, blue = fruits, green = numbers, orange = body, purple = colors, brown = transportation, gray = random words)

4 Experimental Results

4.1 Experimental Setup

Data: ZWJCYLLK3.0 (Chinese corpus), 3 shards (1/3/5), totaling 11,832,314 samples; fixed sequence length 256; vocabulary size 80,000 (sentencepiece BPE). Optimizer: Adam (lr=0.001), batch=8, 10,000 steps. Hardware: Tesla V100-PCIE-32GB, PyTorch 2.8.0+cu128.

Evaluation: For each `[MASK]` position, top-1 is selected from 100 candidates (99 random + 1 target), and the average acc/rank is computed. Explanation of the 100-candidate setting: under an 80k vocabulary, full-vocabulary top-1 is penalized by sparsity (random baseline 1/80000); the 100-candidate setting measures relative discriminability, consistent with related work.

4.2 Three Training Objectives

Key observation: Inverse-causal zero-shot generalization (0.405) is higher than specialized training (0.315). The forward-causal model (10,000 steps) has already learned word–position mappings; reverse inference only requires vocabulary priors + position reversal. Specialized training learns the inverse direction from scratch, and 10,000 steps are insufficient to surpass the prior. This indicates that what the model learns is a direction-independent wave mapping.

Table 2: Ratio of within-class distance to random distance for each category

Category	Within-class distance / Random distance
Animals	0.909
Fruits	0.919
Mathematics	0.777
Body parts	0.902
Colors	0.848
Vehicles	0.861

Figure 2: PhaseFormer (2L128D) training loss over 10,000 steps (left) and MLM 100-candidate accuracy (right, final 0.435)

4.3 Comparison with the Standard Transformer

PhaseFormer reaches 84.5% of the performance of a standard Transformer (0.435/0.515).

5 Related Work

Existing work (iFairy [4], Fairy2i [5], qFHRR [6]) treats phase as a compression tool: they use discretely quantized phase (2-bit) to approximate real-number operations, and regress to real-number linear combinations in the aggregation step. The essential difference between this paper and them is that we treat phase as a semantic carrier—the three-step core mechanism (gradient update / relation computation / information aggregation) is entirely replaced by phase-based methods, rather than using phase to approximate real numbers. PhaseFormer adopts continuous floating-point angles, replaces dot products with the cosine of phase differences, completes aggregation in the angular domain using circular means, and verifies π-distance complementarity and semantic clustering structure in phase space.

Table 3: Accuracy of PhaseFormer on Three Training Objectives

Objective	acc	rank	top10
MLM (cloze filling)	0.435	0.140	0.690
Causal language model	0.430	0.183	0.630
Reverse causal (specialized training)	0.315	0.174	0.630
Reverse causal (zero-sample generalization)	0.405	0.156	0.635

Table 4: PhaseFormer vs. Standard Transformer (Same Scale)

Model	Steps	acc (100 candidates top-1)
PhaseFormer	10000	0.435
Standard Transformer	10000	0.515

6 Conclusion

PhaseFormer demonstrates that language modeling without real-valued matrix multiplication is feasible:

1. **Phase difference can replace the dot product:** MLM 0.435, causal 0.430, reverse 0.315/0.405, reaching 84.5% of the standard Transformer (0.435/0.515)
2. **There are geometric identities in phase space:** the output rotated by π generates the distance complementarity $d(\phi, w_p[X] + \pi) = \pi - d(\phi, w_p[X])$

3. **Phase-vector encoding organizes semantic information:** six classes of semantic clustering are confirmed (within-class distance < random distance)

Value proposition: We demonstrate the feasibility of a fundamentally different paradigm—using phase differences instead of dot products, using angular updates instead of weight updates, and using wave interference instead of real-valued linear combinations.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019.

[3] T. B. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems* (NeurIPS), 2020.

[4] F. Wang, X. Tan, B. Huang, Y. Zhang, G. Wang, P. Cong, and T. Yang, "iFairy: the First 2-bit Complex LLM with All Parameters in $\{\pm 1, \pm i\}$," arXiv preprint arXiv:2508.05571, 2025.

[5] F. Wang, X. Tan, B. Huang, Y. Zhang, G. Wang, P. Cong, and T. Yang, "Fairy2i: Training Complex LLMs from Real LLMs with All Parameters in $\{\pm 1, \pm i\}$," arXiv preprint arXiv:2512.02901, 2026.

[6] S. Snyder, H. Poursiami, and M. Parsa, "qFHRR: Rethinking Fourier Holographic Reduced Representations through Quantized Phase and Integer Arithmetic," arXiv preprint arXiv:2604.25939, 2026.

[7] J. Nunley, "Attention as Frustrated Synchronization," arXiv preprint arXiv:2606.18694, 2026.

[8] P. Zhang, W. Hui, B. Wang, D. Zhao, D. Song, C. Lioma, and J. G. Simonsen, "Complex-valued Neural Network-based Quantum Language Models," *ACM Transactions on Information Systems*, vol. 40, no. 4, 2022.

[9] Y. Kuramoto, "Self-entrainment of a population of coupled non-linear oscillators," in *International Symposium on Mathematical Problems in Theoretical Physics*, 1975.

Note: Figure translations are in progress. See original paper for figures.