Resonance Transformer: Experimental Findings

Abstract

We present a modified transformer architecture that incorporates phase-based embeddings and resonance-biased attention. In controlled experiments comparing four model configurations on identical data, the resonance architecture achieved 57% lower perplexity than a standard transformer baseline. This report documents our methodology, results, and limitations.

---

1. Experimental Setup

1.1 Data

- **Source**: OpenWebText (5% subset)
- **Training examples**: 200,000 sequences
- **Sequence length**: 256 tokens
- **Total tokens**: ~51M
- **Tokenizer**: GPT-2 (50,257 vocabulary)

1.2 Training Configuration

- **Epochs**: 2
- **Batch size**: 8
- **Gradient accumulation**: 4 steps
- **Optimizer**: AdamW (lr=3e-4, weight decay=0.01)
- **Hardware**: NVIDIA RTX 4070 Ti (16GB VRAM)

1.3 Models Tested

Model	Parameters	Architecture	Phase Initialization
Standard Baseline	123.81M	Standard transformer	None
Model A	125.45M	Resonance transformer	Random
Model B	125.45M	Resonance transformer	Phonetic (CMU Dictionary)
Model C	45M	Resonance transformer	Phonetic (CMU Dictionary)

2. Architecture

2.1 Standard Transformer (Baseline)

Standard causal attention:

```
attention = softmax(QK^T / √d)
```

Configuration: 768 embedding dim, 12 layers, 12 heads, 3072 FFN dim.

2.2 Resonance Transformer (Models A, B, C)

Each token receives two embeddings:

- **Semantic embedding**: Standard learned embedding (768 dim)
- **Phase embedding**: Learned phase vector (32 dim), projected to 768 dim

Final token representation:

```
token = (1 - blend) × semantic + blend × project(phase)
```

Where `blend` is a learnable parameter (initialized at 0.3).

Attention is modified to include a resonance bias:

```
attention = softmax(QK^T / √d + λR)
```

The resonance matrix R is computed as:

```
R[i,j] = mean(cos(phase[i] - phase[j]))
```

This biases attention toward tokens with similar phase vectors.

2.3 Phonetic Initialization

For Models B and C, phase embeddings were initialized using the CMU Pronouncing Dictionary. Tokens sharing the same rhyme signature (stressed vowel + following phonemes) received similar phase vectors.

- Phonetically initialized: 35,431 tokens (70%)
- Randomly initialized: 14,826 tokens (30%)

3. Results

3.1 Final Perplexity

Model	Parameters	Val Perplexity	vs Baseline
Standard Baseline	123.81M	148.90	—
Model A (Resonance, Random)	125.45M	63.88	-57.1%
Model B (Resonance, Phonetic)	125.45M	63.17	-57.6%
Model C (Resonance, Phonetic, Small)	45M	70.92	-52.4%

3.2 Training Progression

Epoch 1 Validation Perplexity:

Model	Epoch 1 PPL
Standard Baseline	261.68
Model A	87.58
Model B	85.67
Model C	Not recorded separately

Epoch 2 Validation Perplexity:

Model	Epoch 2 PPL
Standard Baseline	148.90
Model A	63.88
Model B	63.17
Model C	70.92

3.3 Learned Parameters

The resonance models learned the following blend values:

Model	Embedding Blend	Interpretation
Model A	0.410	41% phase, 59% semantic
Model B	0.467	47% phase, 53% semantic

Models chose to use the phase information substantially, suggesting it provides useful signal.

---

4. Analysis

4.1 Resonance Architecture Effect

The primary finding is the large gap between the standard transformer and the resonance models:

- Standard: 148.90 PPL
- Resonance (random init): 63.88 PPL
- **Improvement: 57%**

This suggests the resonance mechanism itself—biasing attention with phase similarity—provides substantial benefit, independent of how phases are initialized.

4.2 Phonetic Initialization Effect

Comparing Models A and B (same architecture, different initialization):

- Random init: 63.88 PPL
- Phonetic init: 63.17 PPL
- **Improvement: 1.1%**

Phonetic initialization provides a small additional benefit.

4.3 Parameter Efficiency

Model C tested whether phonetic initialization enables compression:

- Model A (126M, random): 63.88 PPL
- Model C (45M, phonetic): 70.92 PPL

Model C, with 36% of the parameters, achieved perplexity within 11% of Model A. For comparison, the standard baseline with 124M parameters achieved 148.90 PPL—more than twice Model C's perplexity despite having nearly 3x the parameters.

5. Limitations

5.1 Data Scale

Our token-to-parameter ratio was approximately 0.4:1. Standard practice suggests 20:1 or higher for well-trained models. All models were undertrained, which may affect how results generalize to full-scale training.

5.2 Single Dataset

Experiments used only OpenWebText. Results may not generalize to other domains or data distributions.

5.3 No Downstream Evaluation

We measured perplexity only. Performance on downstream tasks (question answering, summarization, etc.) was not evaluated.

5.4 Limited Scale

The largest model tested was 126M parameters. Behavior at 1B+ parameters is unknown.

5.5 Phonetic Coverage

The CMU dictionary covers approximately 70% of GPT-2 tokens. Subword fragments, numbers, and rare words received random initialization.

---

6. Computational Overhead

The resonance architecture adds:

Component	Additional Cost
Phase embeddings	+1.6M parameters (+1.3%)
Resonance matrix computation	$O(S^2 \times F)$ per layer
Total overhead	~5% compute

---

7. Conclusions

1. **Resonance attention substantially outperforms standard attention** in this experimental setting, achieving 57% lower perplexity with equivalent parameters.
2. **Phonetic initialization provides modest additional benefit** (1.1%) over random phase initialization.
3. **The architecture shows favorable parameter efficiency**: a 45M resonance model outperformed a 124M standard model.
4. **Models actively use the phase mechanism**: learned blend parameters stabilized around 41-47%, indicating the model finds phase information useful.

These results suggest resonance attention merits further investigation at larger scales with more training data.

---

8. Future Work

- Train to convergence with appropriate token-to-parameter ratios
- Evaluate on standard benchmarks (WikiText-103, LAMBADA)
- Test at 1B+ parameter scale
- Explore richer phonetic representations (continuous similarity vs. binary grouping)
- Evaluate on downstream tasks

Appendix: Model Configurations

Standard Baseline (123.81M parameters)

```
embed_dim: 768
n_layers: 12
n_heads: 12
ff_dim: 3072
```

Resonance 126M (Models A & B)

```
embed_dim: 768
n_layers: 12
n_heads: 12
ff_dim: 3072
n_frequencies: 32
resonance_blend_init: 0.3
resonance_attn_weight_init: 0.1
```

Resonance 45M (Model C)

```
embed_dim: 512
n_layers: 8
n_heads: 8
ff_dim: 2048
n_frequencies: 32
resonance_blend_init: 0.3
resonance_attn_weight_init: 0.1
```