Adaptive Multi-Layer Recursive Preconditioned Attention: Addressing Attention Collapse Through Dynamic Cross-Layer Memory Integration

Initial Findings & Open Questions

Yasar Mulani* Junior Research Intern, Qkrishi

Tanmay Gangurde Junior Research Intern, Qkrishi

Rushikesh Ubale Quantum Solution Architect, Qkrishi

December 2, 2025

Abstract

Traditional transformers compute attention independently at each layer, discarding valuable cross-layer dependencies that may benefit multi-hop reasoning. We propose **AMRPA (Adaptive Multi-Layer Recursive Preconditioned Attention)**, a novel mechanism enabling selective memory propagation across layers through similarity-based gating. Our preliminary experiments reveal something unexpected: **the mechanism's effectiveness scales dramatically with task complexity**—modest gains on simple reasoning, but substantial improvements on the most challenging multi-hop benchmarks. More intriguingly, our analysis uncovers that *selective forgetting*, not comprehensive memory retention, drives performance. Gate variance emerges as a strong predictor of reasoning capability ($\rho = 0.794$, $p < 0.01$), suggesting the model learns to discriminate between signal and noise. These findings challenge conventional assumptions about how transformers should utilize cross-layer information and raise fascinating questions about the nature of learned reasoning pathways.

1 Introduction: The Attentional Amnesia Problem

1.1 What Transformers Forget

Standard transformer layers compute attention patterns independently:

$$A^{(l)} = \text{softmax}\left(\frac{Q^{(l)}K^{(l)T}}{\sqrt{d_k}}\right), \quad O^{(l)} = A^{(l)}V^{(l)} \tag{1}$$

This *memoryless* design forces deeper layers to rediscover patterns already learned by earlier layers. While this independence enables parallel computation and stable training, it may limit capabilities requiring information aggregation across processing steps—precisely what multi-hop reasoning demands.

1.2 Why Existing Cross-Layer Approaches Fall Short

Recent work (Universal Transformers, TransformerFAM, Recurrent Memory Transformers) explored cross-layer dependencies but encountered three fundamental challenges:

Challenge 1: Attention Collapse — Historical patterns dominate, preventing new learning
Challenge 2: Limited Expressiveness — Uniform or single-layer memory access lacks flexibility
Challenge 3: Gradient Instability — Recursive structures create training difficulties

The Missing Piece: No existing method combines *content-aware selective access* with mechanisms to prevent collapse while maintaining trainability.

2 AMRPA: Similarity-Guided Cross-Layer Memory

2.1 Core Design Philosophy

AMRPA rests on a simple hypothesis: **effective reasoning requires selective forgetting, not comprehensive retention**. The mechanism should learn what to remember and what to discard based on current computational context.

*Corresponding author: mulaniyhofficial@gmail.com

2.2 Mathematical Framework

2.2.1 1. Decay-Regularized Memory Construction

Past attention patterns decay exponentially to prevent gradient instability:

$$\tilde{A}^{(l-k)} = \gamma^k \cdot A^{(l-k)} + \epsilon \cdot \text{Uniform}(n, n), \quad \gamma \in [0.5, 0.9], \epsilon = 10^{-3} \tag{2}$$

The decay factor γ ensures exponentially decreasing influence, while noise prevents complete information loss.

2.2.2 2. Dynamic Pattern Selection

An MLP computes context-dependent weights for historical patterns:

$$\alpha_k^{(l)} = \text{softmax}\left(\text{MLP}_\alpha\left([Q^{(l)}, \text{proj}(\tilde{A}^{(l-k)})]\right)\right) \tag{3}$$

$$M^{(l)} = \sum_{k=1}^{\min(l-1, w^{(l)})} \alpha_k^{(l)} \cdot \tilde{A}^{(l-k)} \tag{4}$$

This allows the network to dynamically emphasize relevant historical patterns based on current queries.

2.2.3 3. Similarity-Based Gating (Key Innovation)

Unlike prior gating mechanisms that concatenate features, we explicitly measure query-memory relevance:

$$\hat{M}^{(l)} = \text{proj}(M^{(l)}), \quad s^{(l)} = \frac{\langle Q^{(l)}, \hat{M}^{(l)} \rangle}{\sqrt{d_k}} \tag{5}$$

$$G^{(l)} = \sigma(\gamma_g \cdot s^{(l)} + b_g) \tag{6}$$

where γ_g and b_g are learnable parameters. The gate directly models semantic alignment between current computation and historical memory.

2.2.4 4. Memory-Augmented Attention

The final attention integrates base and gated memory signals:

$$A^{(l)} = \text{softmax}\left(\frac{Q^{(l)} K^{(l)T}}{\sqrt{d_k}} + G^{(l)} \odot W_{\text{mem}}^{(l)} M^{(l)}\right) \tag{7}$$

2.3 Layer-Adaptive Memory Depth

Different layers access different memory windows based on computational depth:

$$w^{(l)} = \begin{cases} 1 & \text{if } l \leq 2 \\ \lfloor \log_2(l) \rfloor + 1 & \text{if } 2 < l \leq 8 \\ 4 & \text{if } l > 8 \end{cases} \tag{8}$$

Early layers have limited history; deeper layers access broader context.

3 Preliminary Findings: The Complexity-Gain Relationship

3.1 Experimental Setup

We integrated AMRPA into the final 4 layers of RoBERTa-base and evaluated on three multi-hop question answering benchmarks of increasing difficulty:

HotpotQA: 2-hop reasoning over Wikipedia paragraphs
2WikiMultihop: 2-3 hop reasoning with entity disambiguation
MuSiQue: 2-4 hop reasoning with distractor paragraphs (considered hardest)

Dataset	Hops	Baseline		AMRPA		Pattern
		F1	EM	F1	EM	
HotpotQA	2	51.75	39.82	58.66	45.84	Consistent
2WikiMultihop	2-3	66.33	57.67	76.01	67.37	Strong
MuSiQue	2-4	27.82	19.01	47.56	37.07	**Dramatic**

Table 1: Performance across three multi-hop reasoning benchmarks (F1 and Exact Match scores). AMRPA's absolute improvements scale with task complexity—MuSiQue, the most challenging benchmark requiring navigation through distractor paragraphs, shows the largest gains on both metrics.

3.2 The Unexpected Pattern

What This Suggests: The pattern is striking—baseline F1 varies dramatically across datasets (51.75 → 66.33 → 27.82), reflecting their difficulty. Yet AMRPA consistently lifts performance, with the most substantial gains on the hardest task (MuSiQue: 27.82 → 47.56 F1, 19.01 → 37.07 EM). This selectivity implies the mechanism specifically targets reasoning bottlenecks rather than providing generic enhancement.

Comparison with Published Baselines: Literature reports RoBERTa-base on these benchmarks typically achieves: HotpotQA F1 52-54%, 2WikiMultihop 65-68%, MuSiQue 25-30%. Our baseline aligns with published results, while AMRPA substantially exceeds them across all tasks and both metrics.

3.3 The Counterintuitive Discovery: Less Memory, Better Performance

We instrumented AMRPA to track internal behavior during training, searching for what drives improvements. The analysis on 2WikiMultihop across 10 training epochs revealed unexpected patterns:

Internal Metric	Correlation (ρ)	p-value	Interpretation
Gate Variance	+0.794	0.0061	Strong positive (highly sig.)
Gate Impact (mean)	-0.770	0.0092	Strong negative (highly sig.)
Alpha Diversity	-0.576	0.0816	Moderate negative (marginal)
Memory Contribution	+0.394	0.2600	Weak positive (not sig.)

Table 2: Mechanism analysis reveals surprising relationships. Gate selectivity (variance) strongly predicts performance, while average gate usage shows inverse relationship.

Mechanism Behavior Across Benchmarks:

Dataset	Gate Impact	Gate Variance	Alpha Diversity	Memory Contrib.
HotpotQA	0.635	0.028	0.042	8.852
MuSiQue	0.451	0.033	0.208	3.715
2WikiMultihop	0.442	0.018	0.150	5.616

Table 3: Internal metrics across datasets. Notice: harder task (MuSiQue) shows lower gate impact but higher variance—selective, not uniform, memory usage.

3.3.1 Finding #1: Selective Gating Predicts Success

Gates with *higher variance* produce better reasoning, even when average activation is *lower*. This challenges the intuition that "more memory equals better performance."

Interpretation: The model learns discriminative gating—strongly accepting crucial patterns while aggressively filtering noise. Performance stems from knowing what to forget, not from remembering everything.

3.3.2 Finding #2: Two-Phase Learning Dynamics

During training, we observe distinct phases:

Phase 1 (Early Training): Rapid performance gains with gate exploration — the model discovers *what* to gate

Phase 2 (Late Training): Performance plateau while gates continue refining — the model optimizes *how* to gate

This suggests a form of meta-learning: AMRPA first learns task structure, then optimizes information routing. The mechanism isn't just adding capacity; it's learning a reasoning pathway.

3.4 Theoretical Implications

Why Complexity Helps AMRPA: Harder problems require more selective information filtering. Simple tasks can succeed with local attention; complex multi-hop reasoning benefits from intelligent cross-layer routing.

Attention Collapse Prevention: The similarity-based gating $G^{(l)}$ bounds memory influence by semantic relevance:

$$\|G^{(l)} \odot W_{\text{mem}}^{(l)} M^{(l)}\|_2 \leq \sigma_{\max}(W_{\text{mem}}^{(l)}) \cdot \|M^{(l)}\|_2 \cdot \sigma(\gamma_g \cdot \cos(Q^{(l)}, \hat{M}^{(l)}) + b_g) \tag{9}$$

Irrelevant memory (low similarity) contributes minimally, preventing outdated patterns from dominating.

4 Open Questions & Future Directions

Our preliminary findings raise intriguing questions:

Scaling: Does AMRPA transfer to GPT-style generation? How do gains scale with model size (350M, 1B, 7B)? What about mathematical reasoning, long-context modeling, code understanding?

Mechanism: What patterns do gates select (bridge entities, keywords, relations)? Can we predict gate behavior from task structure? Why does variance correlate so strongly ($\rho = 0.794$)?

Architecture: Could $\gamma^{(l)}$ be learned per layer/task? What about sparse memory (attention over attention patterns)? Multimodal cross-modal memory?

Theory: Convergence guarantees? Representational capacity gains? Inductive biases characterization?

5 Opportunities for Exploration

The mechanism's complexity-scaling behavior and "less is more" gate pattern suggest deeper principles worth understanding. Natural extensions include:

Validation: Multiple seeds with significance tests, comprehensive baselines (Transformer-XL, Longformer, graph methods), broader tasks

Interpretability: Attention flow visualization, reasoning chain case studies, per-question-type analysis

Extensions: Learnable decay, sparse integration, multimodal applications, efficient attention combinations

Theory: Convergence proofs, complexity analysis, capacity bounds

6 Conclusion

AMRPA demonstrates that similarity-guided selective memory can enhance multi-hop reasoning, with effectiveness scaling proportionally to task complexity. The mechanism learns discriminative gating (high variance, low mean) that filters information rather than accumulating it comprehensively—a "quality over quantity" approach to cross-layer dependencies.

The strong correlation between gate selectivity and reasoning capability ($\rho = 0.794$, $p < 0.01$) suggests this isn't merely an engineering trick but may reflect something fundamental about how transformers can be extended for complex reasoning.

Many questions remain open. We hope these preliminary findings spark interest in the broader community to explore, validate, challenge, or extend these ideas.