Memorial: Lightweight Control Supplement

# 1. Introduction to Lightweight Alternatives

While reinforcement learning (RL) provides a powerful mechanism for learning memory modulation policies in response to emotional context, it can be computationally intensive and slow to adapt in real-time settings.

This supplement outlines lightweight alternatives for modulating the Q (Query), K (Key), and V (Value) components of attention mechanisms in transformer-based architectures, allowing for faster, more efficient, and interpretable memory systems.

# 2. Query Modulation Alternatives (Q)

The query vector represents the emotional and intentional state of the user. Alternatives to RL for Q include:

- Multi-Armed Bandits (UCB, Thompson Sampling): Select from a set of exploration strategies based on estimated emotional reward.  
 - Gating Networks: Use emotion vectors as inputs to a softmax-based gating layer that mixes multiple query strategies.  
 - Hebbian Plasticity: Strengthen query-response paths that were emotionally rewarding through associative learning.

# 3. Key Filtering Strategies (K)

Key vectors represent memory slots or past interactions. Alternatives to RL for K include:

- Associative Memory Models (e.g., Hopfield): Automatically activate memories with high emotional similarity.  
 - Sparse Attention with Learned Masking: Only top-k keys with high emotional alignment are considered.  
 - Entropy-Based Filtering: Suppress or amplify key activations based on the uncertainty or clarity of emotional match.

# 4. Value Selection Methods (V)

Value vectors define how the system responds once a memory is selected. RL alternatives for V include:

- Mixture-of-Experts (MoE): Multiple output modules (empathy, information, reflection), weighted by emotional relevance.  
 - Utility-Based Decision Layer: Simple rules to pick responses with highest emotional utility.  
 - Emotion-Embedding Matching: Match current emotional vector with stylistic templates to select tone and form.

# 5. Integrated Lightweight Framework

By combining the above methods, we create a full lightweight memory architecture:

- Q: Gating network with emotion vector input  
 - K: Sparse key selection with associative memory  
 - V: Mixture-of-Experts for emotionally aligned responses

This system is modular, explainable, and adaptable to real-time interaction constraints.

# 6. Use Cases and Future Potential

These lightweight control alternatives make the emotional memory architecture viable for:

- Mobile or edge AI systems with limited compute  
 - Fast-reacting agents in mental health or education  
 - Transparent AI models where behavior needs to be interpretable

Future work includes hybridizing lightweight control with RL fine-tuning for long-term personalization.