Continual Learning and Memory Models Aksel Kretsinger-Walters, adk2164
Memory Models Overview: Summary adk2164@columbia.edu
COMS 6998: Spring 2026 **Due:** 01/28/2026 11:59 PM ET

Memory Models Overview: Summary

Motivation. Memory mechanisms are central to intelligence, enabling agents to integrate information across time, adapt to new environments, and retain past experiences, and modern neural networks incorporate various forms of memory to support long-range dependencies, context tracking, and continual learning.

Recurrent Memory. Recurrent neural networks (RNNs) maintain memory through hidden states that evolve over time, allowing sequential information to be integrated implicitly, while gated variants such as LSTMs and GRUs mitigate vanishing gradients by controlling information flow through learned gating mechanisms.

Attention-Based Memory. Transformer architectures implement memory via attention mechanisms that explicitly store and retrieve information across tokens, enabling flexible access to past representations, with efficient variants reducing quadratic complexity through sparsity, low-rank approximations, or state compression.

Hybrid Memory Models. Hybrid approaches, including state-space models and structured sequence models, combine recurrence and attention-like mechanisms to achieve scalable long-range memory with improved computational efficiency and inductive bias.

Memory as a System. Rather than a single mechanism, memory in neural systems is best viewed as a collection of interacting components operating at different timescales and capacities, motivating unified frameworks that integrate multiple memory types within a single model.

Complementary Learning Systems. The Complementary Learning Systems (CLS) hypothesis posits fast-learning episodic memory and slow-learning semantic

memory operating in tandem, where interleaved replay enables consolidation of new experiences while preventing interference with existing knowledge.

Episodic Memory. Episodic memory stores individual experiences and is often implemented through replay buffers, context manipulation, or sample retrieval, supporting rapid learning and stabilizing performance in non-stationary environments.

Working Memory. Working memory refers to short-term, task-relevant storage used for computation and reasoning, commonly realized through internal model states, attention mechanisms, or memory-augmented networks trained via meta-learning.

External Memory. External memory systems augment neural networks with explicit, addressable storage structures, such as differentiable memory matrices or retrieval-based databases, enabling scalable storage and flexible access beyond fixed network capacity.

Key Takeaway. Modern memory models emphasize modularity, interaction across timescales, and computational efficiency, with successful systems combining recurrent, attention-based, and replay-driven mechanisms to support learning, reasoning, and adaptation over long horizons.