**Continual Learning and Memory Models** Aksel Kretsinger-Walters, adk2164
**Linear Attention and State Space Models** adk2164@columbia.edu
COMS 6998: Spring 2026 **Due:** 02/04/2026 11:59 PM ET

---

# Linear Attention and State Space Models

**Motivation.** Transformer self-attention scales quadratically with sequence length, motivating alternative formulations that enable long-context modeling with linear memory and constant-time inference.

**Transformers as RNNs.** Linear attention reformulates self-attention as a recurrent update by expressing attention as a kernelized dot product, allowing the model to maintain a fixed-size hidden state that summarizes past context.

**Linearized Attention.** By replacing the softmax kernel with a feature map, attention outputs can be computed using accumulated outer products of keys and values, reducing memory growth from sequence length to hidden-state size.

**Recurrent State Update.** The attention state admits a recursive update rule, enabling constant-time inference by updating the state with each new token rather than recomputing attention over the full history.

**Memory Capacity Limits.** Outer-product memory states have finite rank, implying that beyond a certain number of stored key-value pairs, interference arises similarly to capacity limits in Hopfield networks.

**DeltaNet and Fast Weights.** DeltaNet introduces a learned overwrite rule that selectively replaces existing memories, interpreting memory updates as online least-squares optimization and linking linear attention to fast weight programming.

**Memory as Optimization.** Viewing memory updates as gradient descent enables generalization beyond linear maps, culminating in Test-Time Training (TTT), where the hidden state itself is a learnable model updated online.

**Test-Time Training.** TTT treats the hidden state as neural network parameters optimized during inference via self-supervised objectives, blurring the boundary between memory, learning, and inference.

**State Space Models.** State Space Models compress sequence history into a latent dynamical system, enabling linear-time training and constant-time inference by evolving a hidden state governed by learned transition operators.

**Selective State Spaces (Mamba).** Mamba introduces input-dependent parameterization of state updates, allowing selective forgetting and retention while preserving efficient parallel computation.

**Unified Perspective.** Modern sequence models form a hierarchy from attention-based lookup to dynamic compression and online learning, unifying transformers, fast-weight memories, and state space models under a common recurrent framework.