TECHNISCHE UNIVERSITÄT DARMSTADT

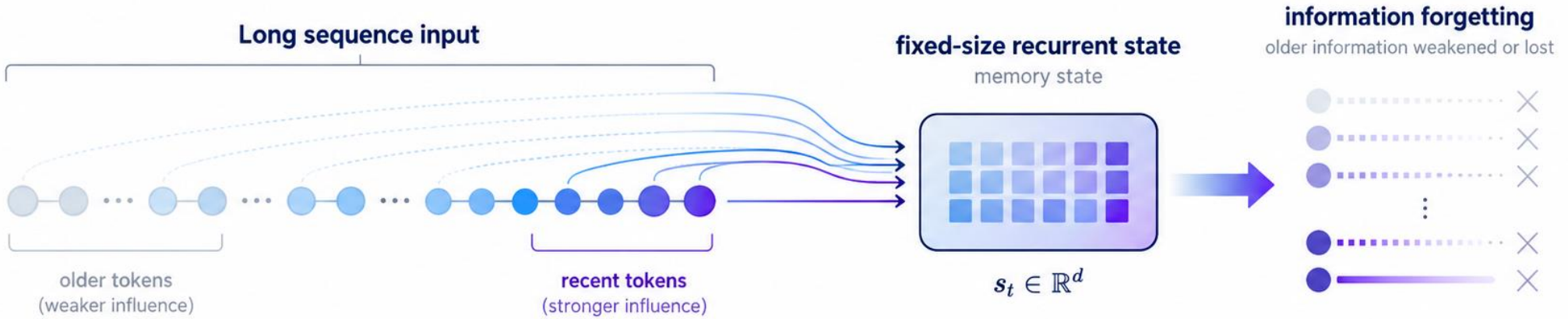
EXPLORING THE FORGETTING PROBLEM IN LINEAR ATTENTION

M.SC JINGCUN WANG

MOTIVATION

- Linear attention enables efficient long-context modeling, but how much past information it actually retains remains unclear.
- Its fixed-size recurrent state introduces a potential efficiency–memory trade-off, where new information may overwrite old information.
- Such forgetting may become a critical bottleneck for tasks requiring long-range dependency and persistent memory.
- Understanding the forgetting mechanism can help us diagnose current limitations and design better linear-attention architectures.

Forgetting in Linear Attention



1. Long context
many tokens over time

2. State compression
compress history into fixed size state

3. Information forgetting
older signals are weakened or lost

YOUR TASK

- Quantify how much information is preserved over time.
- Measure how information retention changes with token distance.
- Build a retention curve for analyzing long-term memory.

- Request:
  - Basic Python programming
  - Basic understanding of Transformers
  - Basic knowledge of linear algebraIdea
- Prerequisites
  - Familiarity with Linear Attention
  - Understanding of recurrent/state-based computation