NOTE: This is a draft for **Stanford CS**. Compile with `final` option for camera-ready copy.

Drawing linear-recurrence to the limit: *Infinite* context modeling

Tushaar Gangavarapu (TG352@cornell.edu)

The motivation for linear-time models may be less compelling today, given the efficiency advancements in Transformers. Are Transformers the endgame?[1] I think not. Are linear models? I'm unsure!

As we become increasingly hardware-aware in advancing Transformers—through techniques like Flash-$*$ algorithms (Fu et al., 2023; Shah et al., 2024) and compute-optimal training (Muennighoff et al., 2023)—the previously prohibitive quadratic complexity of Transformers has become less of a limitation for pretraining. The real "brick wall," however, lies in high inference costs, especially for the latest Transformers like OpenAI o1, which may be employing implicit chain-of-thought (CoT) with stepwise internalization. Even with speculative decoding and KV-cache optimizations, relying on attention for very large contexts (e.g., 100M tokens[2]) is infeasible.

Distilling pretrained Transformers to linear-recurrence models (Zhang et al., 2024b; Wang et al., 2024b; Zhang et al., 2024a; Bick et al., 2024) seems promising. However, it is uncertain if these distilled models retain the weaknesses of their base models, particularly in handling long contexts. Moreover, the expressiveness of linear-recurrence models is limited by their fixed-size hidden memory Jelassi et al. (2024). The recently proposed Based model (Arora et al., 2024) builds on prior efforts by combining linear attention for larger hidden states with sliding-window attention for associative recall. However, further evaluation is needed to assess its performance in long-context retrieval tasks (e.g., needle-in-a-haystack, hash-hop[3], etc.). In our experiments with needle-in-a-haystack evaluations, moderately-sized models, such as RecurrentGemma-9B (with linear recurrence using a hidden state of $4,096$, and a local attention window of $2,048$ tokens), often struggle beyond $4\times$ the local attention window. To this end, we need to devise mechanisms that reduce token storage redundancy (unlike Transformers) that support dynamic memory scaling.

As a part of my PhD at Stanford, I plan to focus on:

(a) analyzing the Pareto frontier of the context length vs. associative recall across varying hidden-state capacities, and

(b) developing models to push this frontier to achieve efficient and reliable long-context modeling.

Designing new architectures will require understanding the limitations of current models. For instance, in hybrid models that combine attention and recurrence, it is essential to explore what long-context information persists across local attention boundaries. How would this change if the model could anticipate specific information needs in advance?[4]

Given my research goal of advancing long-context modeling architectures, I am particularly interested in working with *Prof. Chris Ré, Prof. Tengyu Ma, and Prof. Carlos Guestrin* at Stanford CS. Prof. Chris Ré's lab extensively explores alternate-attention architectures that are inherently compute-efficient (e.g., S4 models, Morach mixer, Based) and hardware-aware (e.g., Flast-$*$, ThunderKittens), which aligns closely with my focus on developing scalable, efficient long-context models. I am also interested in Prof. Tengyu Ma's work on understanding how Transformers enable in-context learning, as this could be invaluable for designing recurrent models that manage information effectively across attention windows. Prof. Carlos Guestrin's research on self-supervised learning via test-time training enhances the expressivity of hidden states in recurrent models, aligning with my objectives of improving recall capacity in linear-time models.

I am drawn to these problems based on my research experiences, which are summarized below.

[1] Sasha's "Transformers in 2027" bet: `https://www.isattentionallyouneed.com`.

[2] LTM-2-mini handles 100M context windows through a proprietary *sequence-dimension algorithm* that may not rely entirely on attention: `https://magic.dev/blog/100m-token-context-windows`.

[3] HashHop long-context evaluation: `https://github.com/magicproduct/hash-hop`.

[4] This is akin to adding a "remember this" to the prompt, instructing the model to retain specific information.

MambaByte. Over the past year, I worked on MambaByte (Wang et al., 2024a), an application of the Mamba recurrence (Gu and Dao, 2024) to byte sequences, under Prof. Sasha Rush. Byte-level modeling is robust to noise and favorable in multilingual and multimodal settings, but computationally expensive due to longer sequence lengths. To maintain training efficiency without input compression, we used the Mamba model, which evolves a large, fixed-size hidden state. Despite training improvements, byte-level decoding remains slower. To mitigate this, we adapted speculative decoding (Leviathan et al., 2023), using a fast subword model for drafting, followed by byte-level verification via parallel scan. To reduce memory I/O overhead, Mamba avoids materializing hidden states in HBM, preventing restarts from corrected bytes. We extended the CUDA kernel to run the parallel scan twice: once to identify the correction point, and again to extract the hidden state before correction. This allowed MambaByte to achieve speedups comparable to the Mamba subword model.

Performance comparison across models also required care—patching models use fewer FLOPs per byte, while having significantly more parameters. To this end, we benchmarked MambaByte in both parameter-matched and compute-matched settings. Our research: (1) demonstrates the viability of token-free models with subword-like inference speedups, and (2) reaffirms that recurrent models excel in tasks where storing all tokens is unnecessary and information compression is more beneficial.

Recurrence in hybrid models. In this work with Hendrik Strobelt and Prof. Sasha Rush, I am investigating the role of recurrence in hybrid models. Recent advances in mechanistic interpretability focus on distilling model activations into "clean" features. Building on this, we employ k-sparse autoencoders to distill hidden states in hybrid models—specifically RecurrentGemma-9B—into sparse, potentially one-dimensional, δ-orthogonal features. By expanding hidden states into (interpretable) features, we aim to explore how recurrence processes contextual information over time and whether it captures meaningful long-range dependencies. These insights could inform the design of future models that better integrate recurrence and attention for improved long-context modeling.

Conversational forecasting. In my research with Prof. Cristian Danescu-Niculescu-Mizil, I investigate conversational forecasting: predicting when a conversation may escalate into a personal attack. This task is challenging as it requires understanding conversational dynamics rather than isolated utterances. The multi-turn nature demands efficient modeling of lengthy contexts—a limitation for standard Transformers. Additionally, real-time predictions with each new reply require minimizing repeated computations as the context grows.

Our experiments show that large generative models (e.g., GPT-4o) struggle with this task, even with full context, while smaller recurrent and bidirectional models perform more effectively, demonstrating their strength in continuous context analysis. We also examined: (1) Simulating future conversational paths: We found low variability in simulations, indicating that models struggle to capture diverse outcomes. (2) Attention to understand model behavior: Attention often failed to provide meaningful explanations, especially as the context length increased. These findings highlight the need for models that can better handle long-range context and conversational dynamics.

Alongside my research, I am deeply passionate about teaching, as it allows me to inspire and motivate students to explore and advance the field. I was awarded the Cornell Bowers CIS TA Award twice for my contributions to CS 4740 (Natural Language Processing)[5] under Prof. Lillian Lee and CS 4300 (Language and Information) under Prof. Mizil. I am excited to co-instruct CS 3780 (Intro to Machine Learning) at Cornell in Spring 2025. I am eager to grow as an academic by teaching at Stanford.

So, the endgame? LLM research has seen steady progress with efficient algorithms, improved hardware, and compute-optimal scaling. Now, CoT reasoning with search stands as the "strawberry" on top. I believe long-context modeling is crucial in extending the limits of context understanding, and I aim to contribute to this during my PhD.

[5] I am proud of my lecture notes on backpropagation: `https://tinyurl.com/backprop-lec-notes`, and Seagull, an LLM that *attempts* to understand humor: `https://tinyurl.com/seagull-lm`.

Acknowledgments. Inspirations for the research directions in this statement stem in part from extensive discussions on the limitations of MambaByte with Junxiong Wang. Additionally, I thank Sankaraleengam Alagapan, Yann Hicke, Barry Wang, Wenting Zhao, Ramya Lakshminarasimhan, and Katherine O'Connor for their helpful comments on the statement. (The ideas in this statement are my own and have not been created or modified using generative AI tools such as ChatGPT.)

References

S. Arora, S. Eyuboglu, M. Zhang, A. Timalsina, S. Alberti, D. Zinsley, J. Zou, A. Rudra, and C. Ré. Simple linear attention language models balance the recall-throughput tradeoff, Feb. 2024. URL http://arxiv.org/abs/2402.18668. arXiv:2402.18668.

A. Bick, K. Y. Li, E. P. Xing, J. Z. Kolter, and A. Gu. Transformers to SSMs: Distilling Quadratic Knowledge to Subquadratic Models, Aug. 2024. URL http://arxiv.org/abs/2408.10189. arXiv:2408.10189.

D. Y. Fu, H. Kumbong, E. Nguyen, and C. Ré. FlashFFTConv: Efficient Convolutions for Long Sequences with Tensor Cores, Nov. 2023. URL http://arxiv.org/abs/2311.05908. arXiv:2311.05908.

A. Gu and T. Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, May 2024. URL http://arxiv.org/abs/2312.00752. arXiv:2312.00752.

S. Jelassi, D. Brandfonbrener, S. M. Kakade, and E. Malach. Repeat After Me: Transformers are Better than State Space Models at Copying, June 2024. URL http://arxiv.org/abs/2402.01032. arXiv:2402.01032.

Y. Leviathan, M. Kalman, and Y. Matias. Fast Inference from Transformers via Speculative Decoding, May 2023. URL http://arxiv.org/abs/2211.17192. arXiv:2211.17192.

N. Muennighoff, A. M. Rush, B. Barak, T. L. Scao, A. Piktus, N. Tazi, S. Pyysalo, T. Wolf, and C. Raffel. Scaling Data-Constrained Language Models, Oct. 2023. URL http://arxiv.org/abs/2305.16264. arXiv:2305.16264.

J. Shah, G. Bikshandi, Y. Zhang, V. Thakkar, P. Ramani, and T. Dao. FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision, July 2024. URL http://arxiv.org/abs/2407.08608. arXiv:2407.08608.

J. Wang, T. Gangavarapu, J. N. Yan, and A. M. Rush. MambaByte: Token-free Selective State Space Model, Aug. 2024a. URL http://arxiv.org/abs/2401.13660. arXiv:2401.13660.

J. Wang, D. Paliotta, A. May, A. M. Rush, and T. Dao. The Mamba in the Llama: Distilling and Accelerating Hybrid Models, Aug. 2024b. URL http://arxiv.org/abs/2408.15237. arXiv:2408.15237.

M. Zhang, S. Arora, R. Chalamala, A. Wu, B. Spector, A. Singhal, K. Ramesh, and C. Ré. LoLCATs: On Low-Rank Linearizing of Large Language Models, Oct. 2024a. URL http://arxiv.org/abs/2410.10254. arXiv:2410.10254.

M. Zhang, K. Bhatia, H. Kumbong, and C. Ré. The Hedgehog & the Porcupine: Expressive Linear Attentions with Softmax Mimicry, Feb. 2024b. URL http://arxiv.org/abs/2402.04347. arXiv:2402.04347.

(Last compiled: 11/10/2024, 6.49pm.)