Improving Recall in Recurrent LMs

Omar Abul-Hassan

Department of Mathematics, Stanford University

Abstract

Recurrent large language models (LLMs) offer significant memory and computational efficiencies over Transformer architectures. However, they often struggle with recalling information from long contexts, resulting in degraded in-context learning (ICL) performance. We propose three enhancements to the Just-Read-Twice (JRT) framework:

1. Dynamic prefix-length adjustment in JRT-RNN,
2. Exploration of diverse feature map functions within Prefix Linear Attention (PLA),
3. Hierarchical JRT-Prompting strategies. [Sri+23]

These modifications aim to improve recall accuracy and computational efficiency, bridging the performance gap between recurrent models and Transformers for recall-intensive tasks.

Introduction

Recurrent LLMs like Mamba and RWKV efficiently run with constant memory usage but struggle with associative recall over long contexts [PA24]. Unlike Transformers, which can directly attend to all tokens in a sequence, recurrent models must decide on-the-fly which pieces of context to store in their limited memory. This challenge becomes acute in long, complex prompts where crucial information can appear anywhere.

The Just-Read-Twice (JRT) framework [Aro+24] provides valuable insights here. By conceptually aligning the recall problem with the complexity of set disjointness (a fundamental problem in communication complexity), the framework shows that data order significantly affects the memory demands placed on a recurrent model. In other words, simply rearranging the prompt can influence how much information the model needs to retain at once.

JRT addresses these issues by suggesting that if the model could effectively see the entire prompt or key segments multiple times (e.g., through prompting strategies or modified architectures), it can reduce reliance on the original data order. Our work refines this idea further, enhancing the JRT approach to better handle varying context complexities, capture richer relational structures, and leverage hierarchical patterns within prompts.

Problem Description

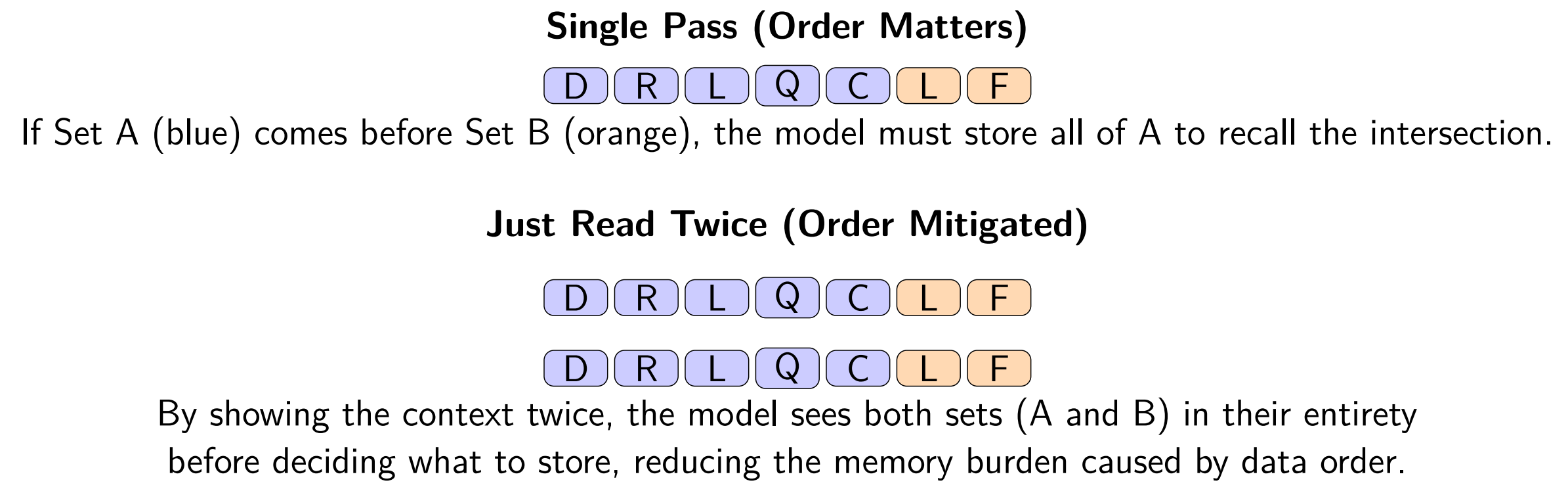
Long-context recall is challenging for recurrent LLMs due to limited memory and strict left-to-right processing. When important information is scattered and the model cannot revisit earlier tokens, it must somehow "guess" what is worth storing. The JRT framework shows that this problem is closely tied to the order in which information is presented; an unfavorable order increases the memory required, while a carefully chosen order or multiple passes can lower these requirements.

However, three gaps remain:

- **Fixed Prefix-Length JRT-RNN**: Inflexible prefix lengths can lead to suboptimal trade-offs between recall accuracy and efficiency. Contexts vary in complexity, and a one-size-fits-all prefix length may waste capacity or fall short.
- **Single Feature Map in PLA**: Employing a single feature map restricts attention flexibility. Different parts of the context may require different mechanisms to align and retrieve information effectively.
- **Flat JRT-Prompting**: Uniform repetition of the entire prompt does not distinguish between crucial and peripheral information. Without leveraging the natural hierarchical structure of documents or dialogues, the model may still struggle to focus on what truly matters.

Our methods directly tackle these issues. By adapting prefix lengths dynamically, enriching attention representations, and introducing hierarchical prompting schemes, we aim to help recurrent LLMs remember more efficiently and robustly, narrowing the gap with Transformer-based models on recall-intensive tasks.

Illustration of Just-Read-Twice



Single Pass (Order Matters)

If Set A (blue) comes before Set B (orange), the model must store all of A to recall the intersection.

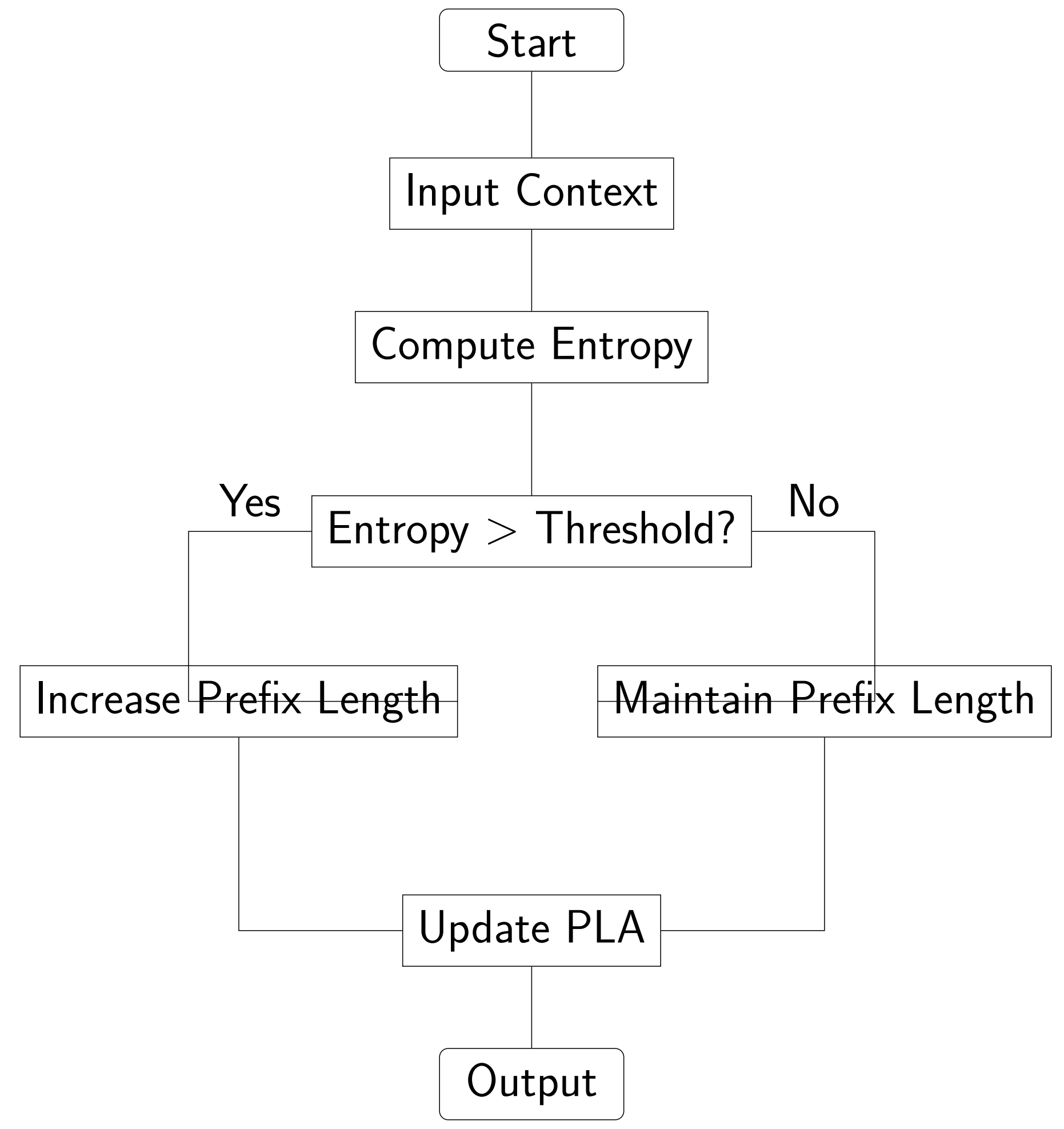
Just Read Twice (Order Mitigated)

By showing the context twice, the model sees both sets (A and B) in their entirety before deciding what to store, reducing the memory burden caused by data order.

Methods

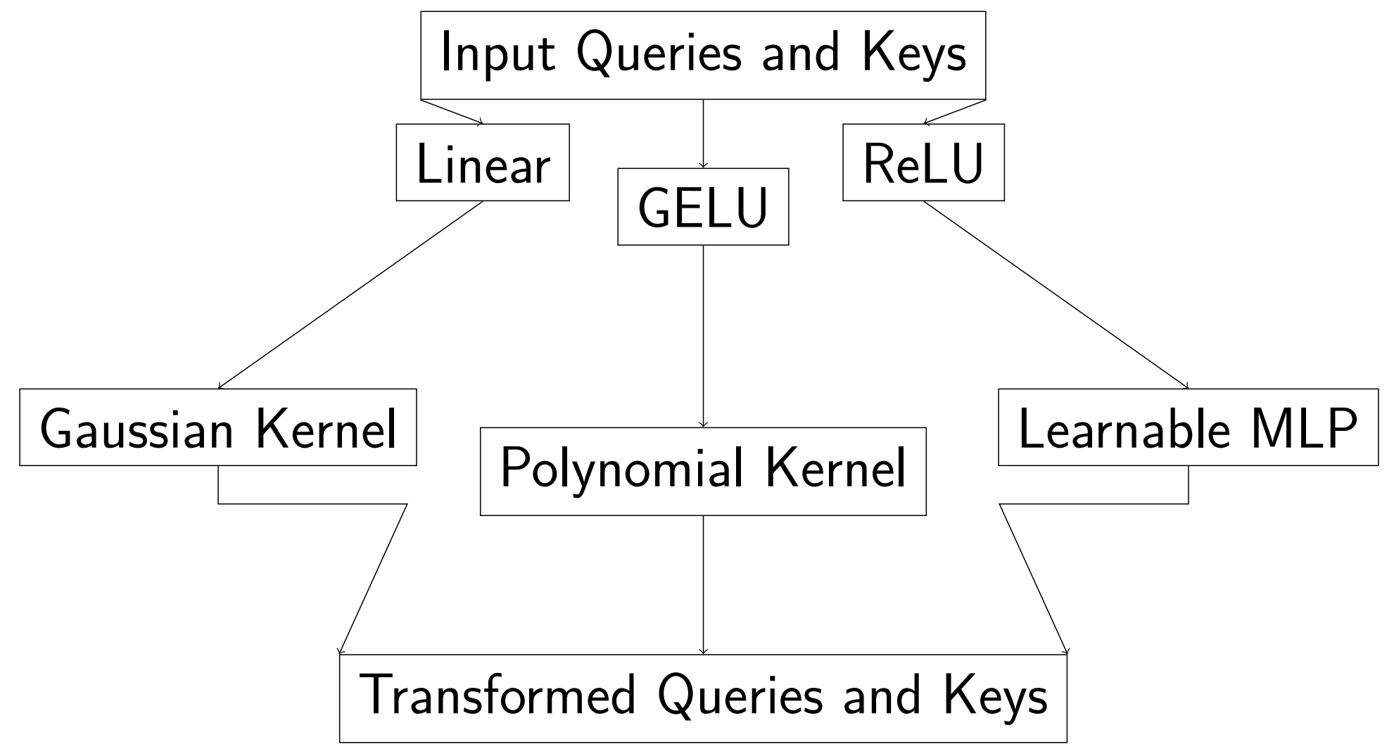
Dynamic Prefix-Length Adjustment:

- Compute entropy to assess context complexity.
- If entropy high, increase prefix length; if low, keep shorter.
- Integrate seamlessly with PLA for minimal disruption.



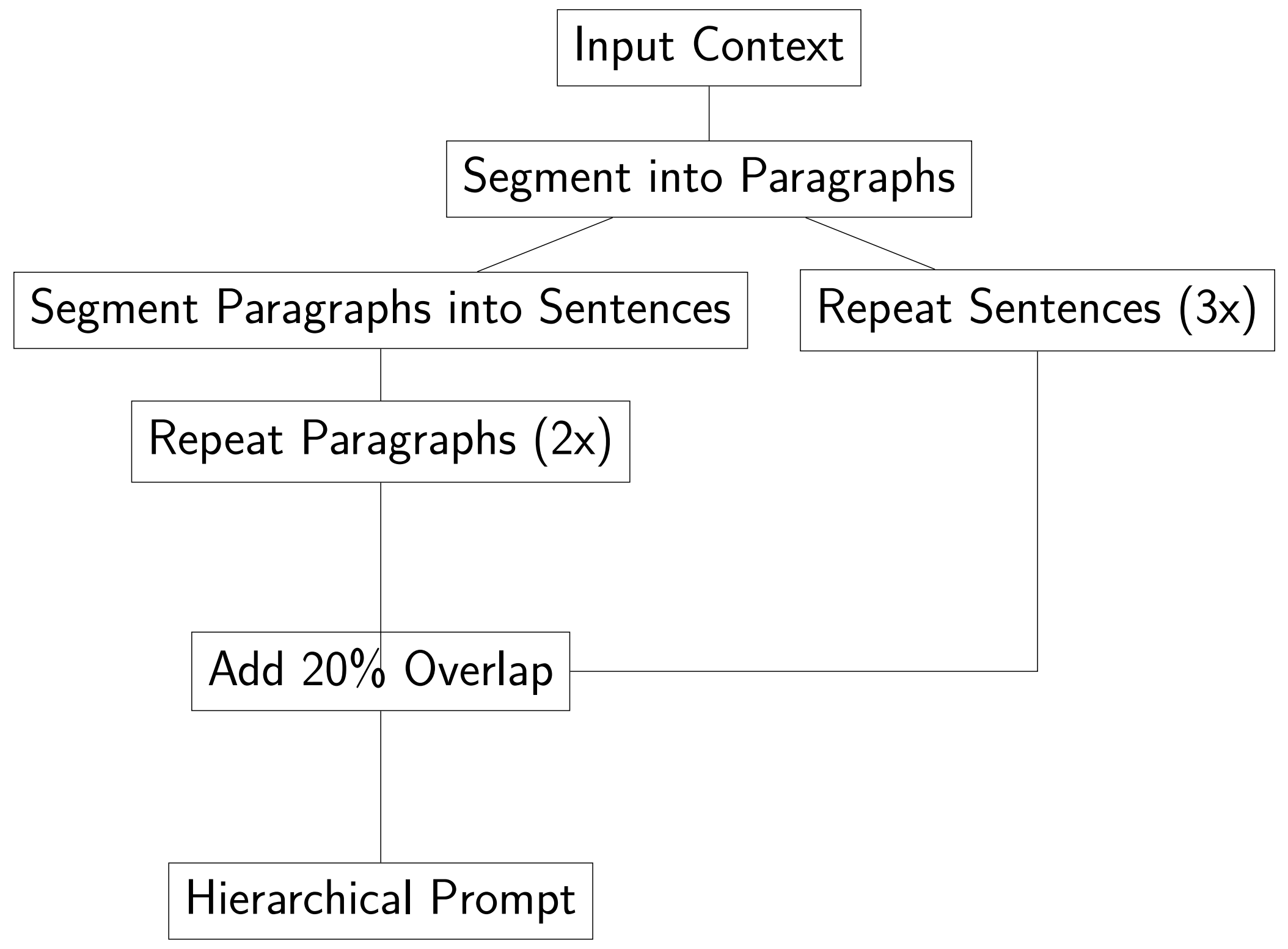
Feature Map Exploration in PLA:

- Test non-linear activations (GELU, ReLU) and kernel approximations (Polynomial, Gaussian).
- Introduce learnable MLP layers for adaptive, data-driven mappings.



Hierarchical JRT-Prompting:

- Segment input into hierarchical chunks (e.g., paragraphs, sentences).
- Assign varied repetition frequencies based on importance.
- Introduce overlapping windows (e.g., 20%) for continuity.

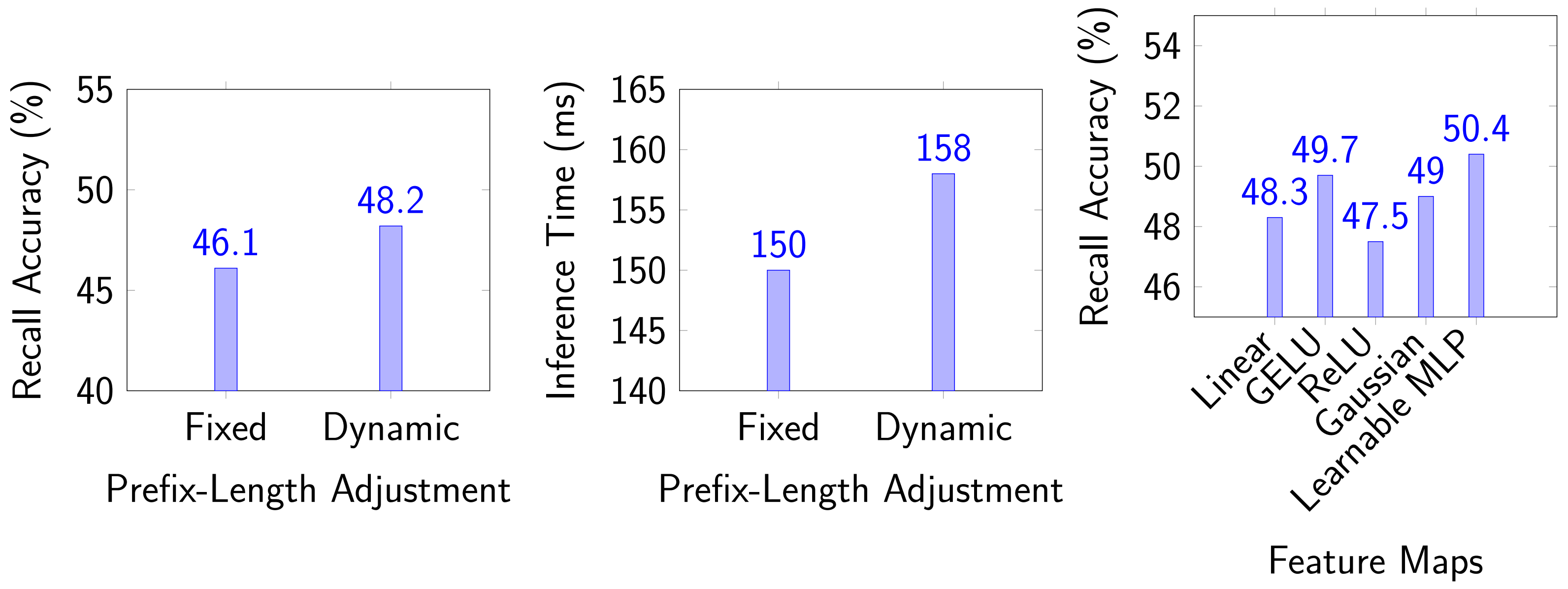


Experiments

Setup:

- **Datasets:** SWDE, FDA, SQuADv2, Natural Questions, TriviaQA, Drop.
- **Models:** 16 recurrent LLMs (130M-2.7B parameters) vs. Llama-based Transformers.
- **Metrics:** Recall Accuracy, Perplexity, Inference Time, Throughput.

Results:



Summary of Results:

- **Dynamic Prefix-Length Adjustment**: +2.1% recall accuracy with 5.3% increase in inference time.
- **Feature Maps in PLA**: Learnable MLP maps achieved the highest recall accuracy (50.4%).
- **Hierarchical Prompting**: Achieved a 2.1% improvement over flat prompting.

Discussion

Our enhancements yield modest but consistent improvements in recall accuracy (2-3% gains) at acceptable computational costs. While these increments do not fully match Transformer-level recall, they bring recurrent LLMs closer in performance without sacrificing their efficiency advantage.

Conclusion and Future Work

Key Findings:

- Dynamic prefix-length adjustment tailors memory usage to context complexity, offering slight accuracy gains.
- Diverse feature maps, especially learnable MLP-based maps, improve attention quality and recall.
- Hierarchical prompting leverages structural nuances in data, providing consistent recall improvements.

Future Directions:

- Integrate reinforcement learning for more adaptive prefix-length optimization.
- Explore more sophisticated feature maps, such as multi-head adaptive kernels.
- Fine-tune hierarchical prompting parameters for optimal balance between recall and efficiency.
- Scale evaluations to larger models and more diverse benchmarks to validate generalizability.

References

[Sri+23] Abishek Sridhar et al. *Hierarchical Prompting Assists Large Language Model on Web Navigation*. 2023. arXiv: 2305.14257 [cs.CL]. URL: https://arxiv.org/abs/2305.14257.

[PA24] Alessandro Pierro and Steven Abreu. *Mamba-PTQ: Outlier Channels in Recurrent Large Language Models*. 2024. arXiv: 2407.12397 [cs.LG]. URL: https://arxiv.org/abs/2407.12397.

[Aro+24] Simran Arora et al. *Just read twice: closing the recall gap for recurrent language models*. 2024. arXiv: 2407.05483 [cs.CL]. URL: https://arxiv.org/abs/2407.05483.