Adapting Hybrid Parallel-Head Large Language Models for Southeast Asia

Kin Meng Ng

Singapore University of Technology and Design
kinmeng_ng@mymail.sutd.edu.sg

Abstract

Large language models (LLMs) have rapidly advanced, but their growing compute demands limit accessibility in under-resourced regions like Southeast Asia (SEA). While hybrid architectures combining Attention and State-Space Models (SSMs) offer efficiency gains, most rely on sequential interleaving, leaving the potential of parallel-head mixing largely under-explored. However, the recent Falcon-H1 family of models has demonstrated that parallel-head hybrid architectures are not only viable, but scalable to state-of-the-art levels. I propose investigating this parallel-head architecture as a foundation for efficient, multilingual SEA LLMs. My short-term goal is to adapt Falcon-H1-1.5B via vocabulary expansion and continuous pretraining, mitigating token fragmentation and enabling low-resource adaptation to 9 SEA languages. In the longer term, I will develop a dynamic token routing mechanism to optimize token-level compute allocation within hybrid layers, aiming to maximize efficiency without sacrificing the expressive power needed for complex multilingual contexts. Evaluation will utilize the SEA-HELM framework to assess whether these parallel-hybrid innovations can democratize access to high-performance AI for SEA communities.

Introduction

I focus on efficient large language model (LLM) architectures, specifically hybrid models that mix state-space model (SSM) heads with Transformer attention heads in parallel within a layer. Hybrid LLM combines the high-resolution recall of attention with the long-range, linear-time summarization of SSMs, offering a promising point on the tradeoff between compute, memory, and modeling power.

Improving architectural efficiency is urgent. Current compute-hungry trajectories concentrate capability behind large corporations and are unsustainable. More efficient model families would reduce costs and enable geographically diverse communities to run powerful models locally. This is critical for Southeast Asia (SEA), where under-resourced populations and uneven connectivity (Sermcheep 2024) make standard large models inaccessible in some regions.

My goal is to (1) evaluate parallel-head hybrid effectiveness for SEA languages using Falcon-H1-1.5B, and (2) investigate dynamic gating mechanisms to further enhance the memory efficiency of these hybrid models.

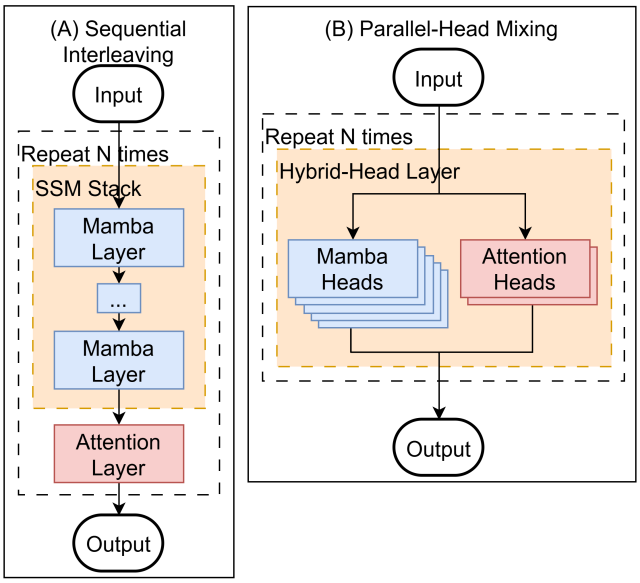

Figure 1: Simplified architectural comparison between (A) Sequential Interleaving and (B) Parallel-Head Mixing. Layer Norm and Feed-Forward Network blocks are omitted.

Background and Related Work

Structured state-space models such as Mamba (Gu and Dao 2024; Dao and Gu 2024), have demonstrated strong performance with linear-time scaling but struggle with precise in-context recall compared to quadratic attention (Gu and Dao 2024; Park et al. 2024; Waleffe et al. 2024). Hybrid designs have rapidly emerged to combine the strengths of both SSMs and attention. As illustrated in Figure 1, two paradigms have emerged: **Sequential Interleaving** alternates layers of attention and SSMs (Glorioso et al. 2024; Lieber et al. 2024; Ren et al. 2024; Blakeman et al. 2025; Team et al. 2025) while **Parallel-Head Mixing**, introduced by Hymba (Dong et al. 2024), and improved upon by Falcon-H1 (Zuo et al. 2025), integrates attention and SSM heads side-by-side.

Parallel mixing intuitively offers a distinct hybrid mod-

eling advantage: it mitigates the representation degradation inherent in sequential stacking hybrids, allowing the attention mechanism to retrieve information from uncompressed inputs rather than the lossy state outputs of preceding SSM layers.

Unfortunately, these hybrid models are largely not trained with SEA languages in mind, even if they are multilingual. For example, Falcon-H1's pre-training data comprised 18 languages, of which none are SEA languages. Crucially, this shift toward efficient hybrid designs has not yet permeated regional modeling efforts. Current SEA languages adaptation (Dou et al. 2024; Ng et al. 2025; Nguyen et al. 2024; Zhang et al. 2025) relies exclusively on full-attention LLMs, leaving a critical gap in understanding how these architectures perform in multilingual, resource-constrained contexts.

Approach and Evaluation

Phase 1 – Multilingual Adaptation

Although Falcon-H1's tokenizer was trained on 121 languages, it still had on average subpar tokenization for SEA languages compared to existing SEA models' tokenizers. Concurrently, standard continual pre-training (CPT) creates a risk of catastrophic forgetting, where the model loses pre-existing knowledge while adapting to new data (Luo et al. 2025).

To extend the tokenizer's vocabulary while mitigating catastrophic forgetting, I employ the Efficient and Effective Vocabulary Expansion (EEVE) method (Kim, Choi, and Jeong 2024), which freezes the backbone to preserve original competencies while integrating new SEA tokens.

I first merge the SEA-LION vocabulary (Ng et al. 2025) with Falcon-H1's. New input embeddings are initialized by averaging their constituent sub-word embeddings to capture semantic approximation, while new output embeddings are initialized using the first sub-word to align with the model's predictive framework. I then deploy a seven-stage CPT framework that prevents catastrophic forgetting by strictly controlling parameter updates. This curriculum moves from training only the newly initialized embeddings and LM heads, to fine-tuning the full model, and finally isolating the hybrid layers for a targeted update step.

This yields a model optimized for SEA languages without compromising the original model's language capabilities and inference efficiency. The resulting model will be evaluated against Sailor2 1B and 3B (Dou et al. 2025), the closest SEA multilingual baselines under 7B parameters, using the SEA-HELM framework (Susanto et al. 2025).

Phase 2 – Architectural Optimization

While Falcon-H1 demonstrates the scalability of parallel-head hybrid architectures, it still applies quadratic attention uniformly to every token, although to fewer heads. This is computationally suboptimal given recent findings from DTRNet (Sharma et al. 2025), which showed that approximately 90% of tokens in a Transformer can bypass the attention mechanism entirely without degrading performance.

Instead of a static design where every token consumes both Attention and SSM compute, I propose a **Context-Preserving Dynamic Router**. In this design, the recurrent Mamba head acts as the "always-on" continuous backbone, preserving the sequence history state h_t for every token. Meanwhile, the computationally expensive Attention head is gated dynamically:

- **Continuous Backbone (Mamba):** Operates on *all* tokens at $\mathcal{O}(N)$. This ensures that even when Attention is skipped, the model maintains a robust, context-aware state representation, unlike the blind linear bypass in DTRNet.
- **Sparse Retrieval (Attention):** Operates at $\mathcal{O}(N^2)$ but is activated only for "retrieval-heavy" tokens identified by the router.

I hypothesize that this "Mamba-backed" routing will allow for even greater attention sparsity than DTRNet, as the high-quality recurrent state should mitigate the information loss from entirely skipping attention.

To implement this, I will employ a token-wise binary router that utilizes a soft routing mechanism during training, computing a weighted sum of execution paths to ensure robust gradient propagation. Simultaneously, I will optimize this router with a cross-entropy loss enhanced with an auxiliary load-balancing loss inspired by DTRNet to prevent routing collapse. This architectural modification aims to deliver even more memory and inference efficiency, by further minimizing how often attention is used at a token level.

Discussion

I expect to find that my model will not only demonstrate effective multilingual adaptation via the EEVE-based CPT pipeline but also reveal distinct efficiency advantages through dynamic token routing. By allowing the model to route high-entropy tokens, such as code-switched boundaries, to attention heads, and syntactic tokens only to the SSM heads, I anticipate a "compute-adaptive" behavior that significantly lowers inference costs for dense SEA languages. Furthermore, analyzing these routing patterns will offer linguistic insights into how the model allocates compute across different token types, such as discourse markers versus reasoning anchors. Validating this dynamic routing on a 1.5B scale will provide a rigorous proof-of-concept for resource-constrained environments, demonstrating that efficient, local deployment is viable without sacrificing the reasoning capabilities of full-attention transformers.

Conclusion

I propose to define the next generation of efficient LLMs for Southeast Asia by investigating hybrid parallel-head architectures. My contributions will be twofold: (1) adapting Falcon-H1-1.5B to SEA languages using vocabulary expansion and continuous pre-training; and (2) developing a dynamic token routing mechanism that optimizes compute by selectively dispatching tokens to an Attention or a Bypass path while maintaining a constant Mamba update state. This project bridges the gap between architectural innovation and linguistic equity, providing a rigorous roadmap to democratize high-performance, locally runnable AI for underrepresented communities.

References

Blakeman, A.; Basant, A.; Khattar, A.; Renduchintala, A.; Bercovich, A.; Ficek, A.; Bjorlin, A.; Taghibakhshi, A.; Deshmukh, A. S.; Mahabaleshwarkar, A. S.; et al. 2025. Nemotron-h: A family of accurate and efficient hybrid mamba-transformer models. *arXiv preprint arXiv:2504.03624*.

Dao, T.; and Gu, A. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

Dong, X.; Fu, Y.; Diao, S.; Byeon, W.; Chen, Z.; Mahabaleshwarkar, A. S.; Liu, S.-Y.; Van Keirsbilck, M.; Chen, M.-H.; Suhara, Y.; et al. 2024. Hymba: A hybrid-head architecture for small language models. *arXiv preprint arXiv:2411.13676*.

Dou, L.; Liu, Q.; Zeng, G.; Guo, J.; Zhou, J.; Mao, X.; Jin, Z.; Lu, W.; and Lin, M. 2024. Sailor: Open language models for south-east asia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 424–435.

Dou, L.; Liu, Q.; Zhou, F.; Chen, C.; Wang, Z.; Jin, Z.; Liu, Z.; Zhu, T.; Du, C.; Yang, P.; et al. 2025. Sailor2: Sailing in South-East Asia with Inclusive Multilingual LLMs. *arXiv preprint arXiv:2502.12982*.

Glorioso, P.; Anthony, Q.; Tokpanov, Y.; Whittington, J.; Pilault, J.; Ibrahim, A.; and Millidge, B. 2024. Zamba: A compact 7b ssm hybrid model. *arXiv preprint arXiv:2405.16712*.

Gu, A.; and Dao, T. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*.

Kim, S.; Choi, S.; and Jeong, M. 2024. Efficient and effective vocabulary expansion towards multilingual large language models. *arXiv preprint arXiv:2402.14714*.

Lieber, O.; Lenz, B.; Bata, H.; Cohen, G.; Osin, J.; Dalmedigos, I.; Safahi, E.; Meirom, S.; Belinkov, Y.; Shalev-Shwartz, S.; et al. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*.

Luo, Y.; Yang, Z.; Meng, F.; Li, Y.; Zhou, J.; and Zhang, Y. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*.

Ng, R.; Nguyen, T. N.; Huang, Y.; Tai, N. C.; Leong, W. Y.; Leong, W. Q.; Yong, X.; Ngui, J. G.; Susanto, Y.; Cheng, N.; et al. 2025. Sea-lion: Southeast asian languages in one network. *arXiv preprint arXiv:2504.05747*.

Nguyen, X.-P.; Zhang, W.; Li, X.; Aljunied, M.; Hu, Z.; Shen, C.; Chia, Y. K.; Li, X.; Wang, J.; Tan, Q.; et al. 2024. SeaLLMs-large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 294–304.

Park, J.; Park, J.; Xiong, Z.; Lee, N.; Cho, J.; Oymak, S.; Lee, K.; and Papailiopoulos, D. 2024. Can mamba learn how to learn? a comparative study on in-context learning tasks. *arXiv preprint arXiv:2402.04248*.

Ren, L.; Liu, Y.; Lu, Y.; Shen, Y.; Liang, C.; and Chen, W. 2024. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *arXiv preprint arXiv:2406.07522*.

Sermcheep, S. 2024. Digital Connectivity In Asean. *Indo-Pacific and ASEAN: New Balances and New Challenges for Asian Integration and Stability*.

Sharma, A.; Najafi, S.; Farinneya, P.; Jamialahmadi, B.; Tahaei, M. S.; Fan, Y.; Rezagholizadeh, M.; Chen, B.; and Jafari, A. 2025. DTRNet: Dynamic Token Routing Network to Reduce Quadratic Costs in Transformers. *arXiv preprint arXiv:2509.00925*.

Susanto, Y.; Hulagadri, A. V.; Montalan, J. R.; Ngui, J. G.; Yong, X.; Leong, W. Q.; Rengarajan, H.; Limkonchotiwat, P.; Mai, Y.; and Tjhi, W. C. 2025. Sea-helm: Southeast asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 12308–12336.

Team, T. H.; Liu, A.; Zhou, B.; Xu, C.; Zhou, C.; Zhang, C.; Xu, C.; Wang, C.; Wu, D.; Wu, D.; et al. 2025. Hunyuan-turbos: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. *arXiv preprint arXiv:2505.15431*.

Waleffe, R.; Byeon, W.; Riach, D.; Norick, B.; Korthikanti, V.; Dao, T.; Gu, A.; Hatamizadeh, A.; Singh, S.; Narayanan, D.; et al. 2024. An empirical study of mamba-based language models. *arXiv preprint arXiv:2406.07887*.

Zhang, W.; Chan, H. P.; Zhao, Y.; Aljunied, M.; Wang, J.; Liu, C.; Deng, Y.; Hu, Z.; Xu, W.; Chia, Y. K.; et al. 2025. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, 96–105.

Zuo, J.; Velikanov, M.; Chahed, I.; Belkada, Y.; Rhayem, D. E.; Kunsch, G.; Hacid, H.; Yous, H.; Farhat, B.; Khadraoui, I.; et al. 2025. Falcon-h1: A family of hybrid-head language models redefining efficiency and performance. *arXiv preprint arXiv:2507.22448*.