Avey-B

Devang Acharya and Mohammad Hammoud

February 2026

avey

Motivation

Inefficiency of Transformers

The quadratic time and memory costs of full self-attention remain a central bottleneck, particularly in bi-directional encoder models

The Avey Architecture

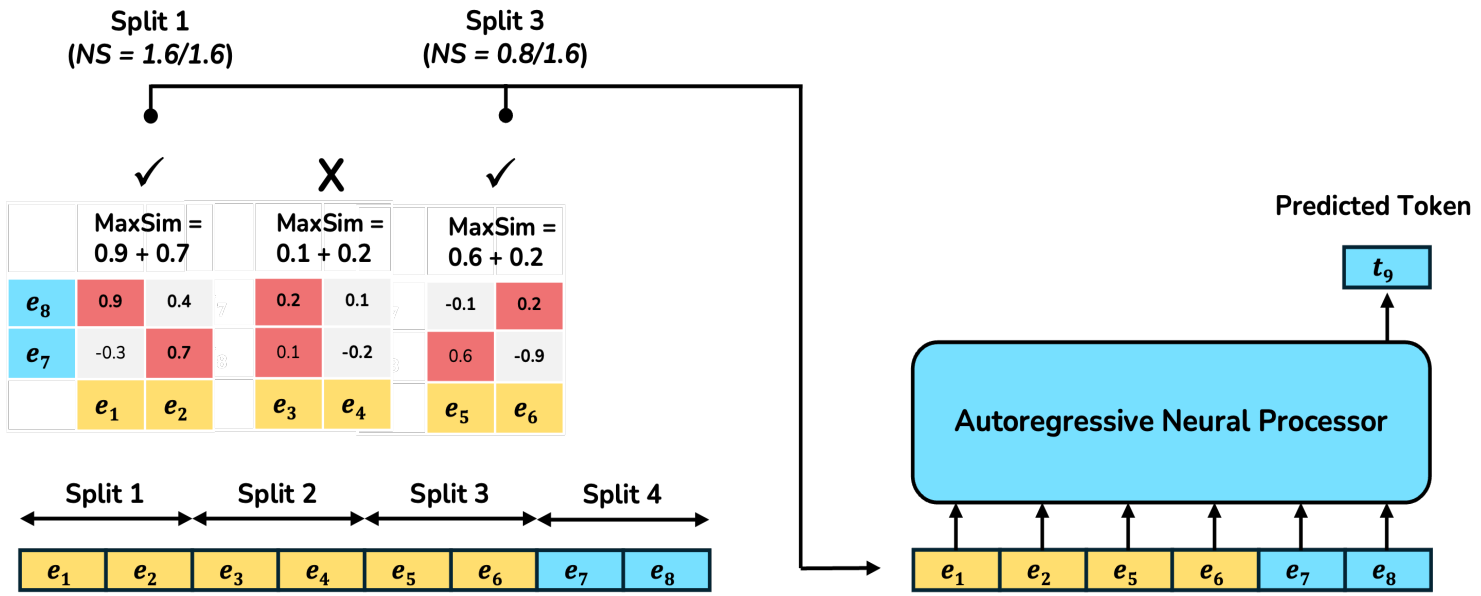
Avey decouples the input sequence length form the model's context width, allowing efficient processing of long sequences

Bi-directional Avey

We reformulate Avey for the encoder-only setting, and introduce several architectural advances improving effectiveness and efficiency

Background: Avey Ranker

- **Sequence Partitioning**: Avey divides the input into S-token splits and ranks preceding splits via MaxSim
- **Weighted Selection**: It then concatenates the current split with the top-*k* splits weighted by normalized MaxSim scores
- **Neural Processing**: Finally, it passes the resulting concatenated representations to a neural processor for contextualization



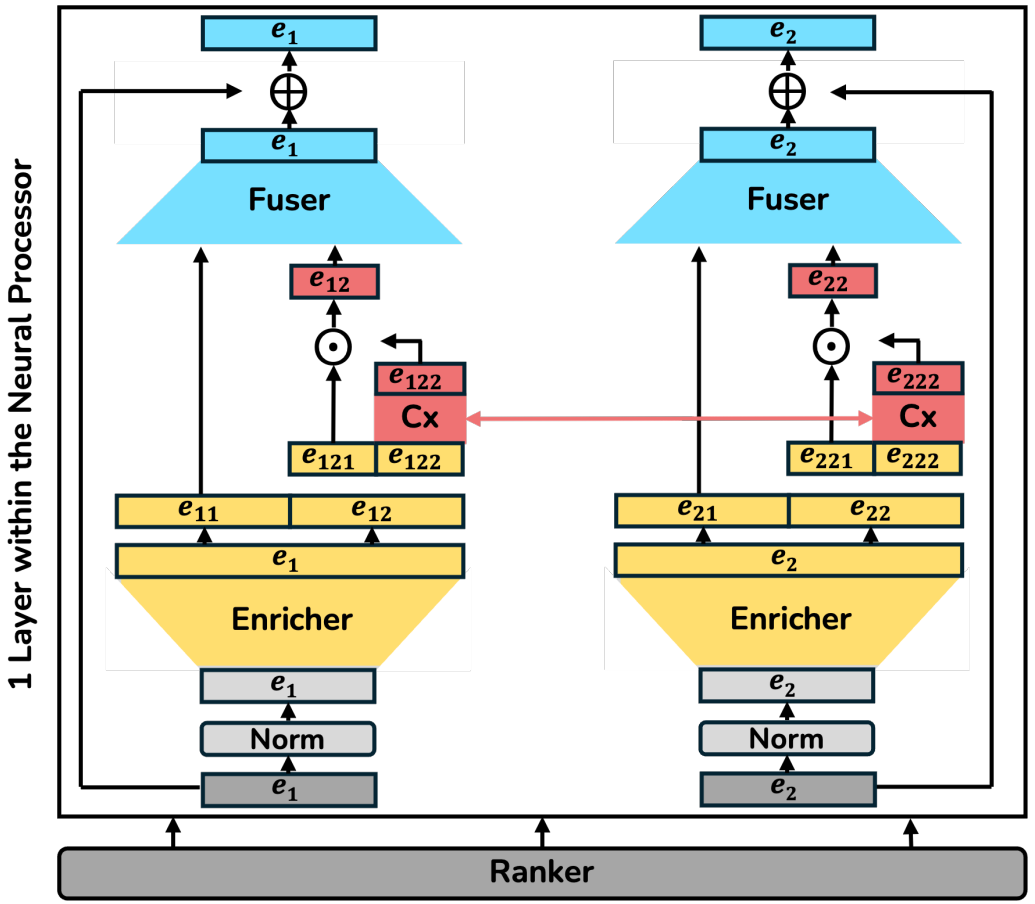
avey

Background: Avey Neural Processor

- **Enricher**: A single-layer, *position-wise* network applied per token, with part of its output bypassed directly to the fuser

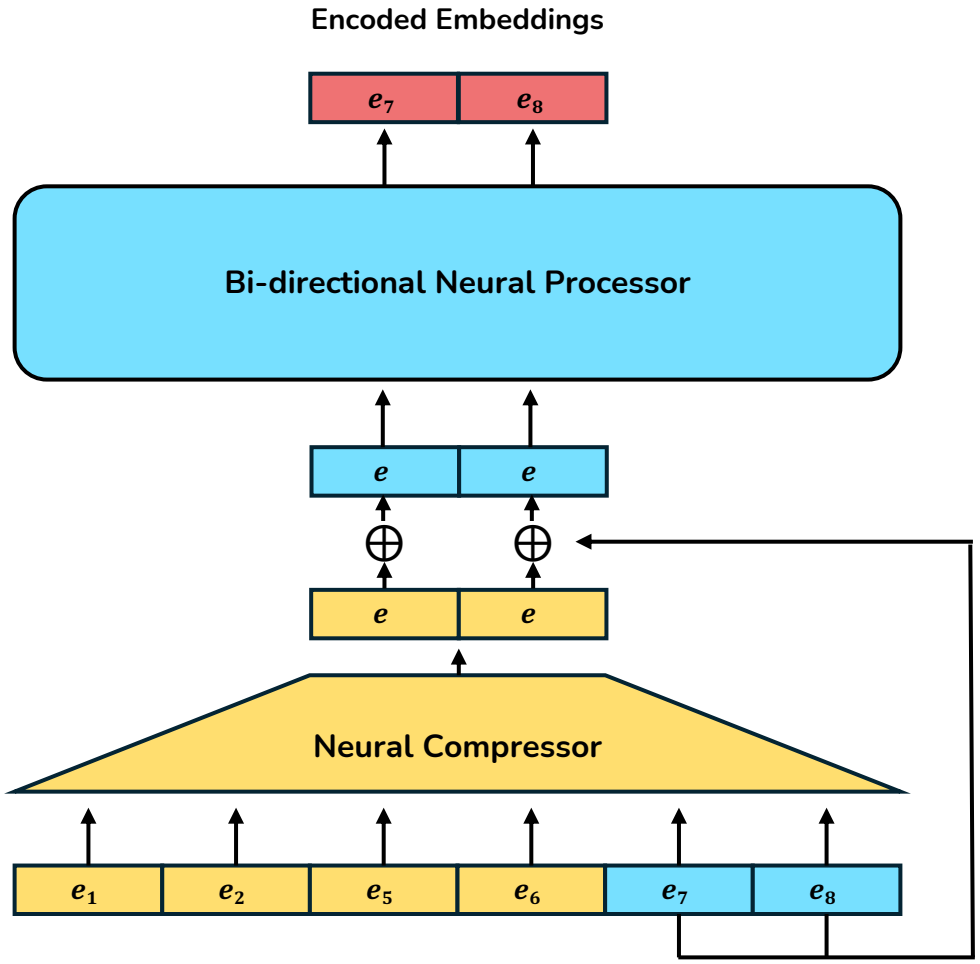
- **Contextualizer**: A single-layer, *embedding-wise* neural network that computes a weighted sum of embeddings

- **Fuser**: A *pointwise* projection that maps bypassed and contextualized features back to model dimension



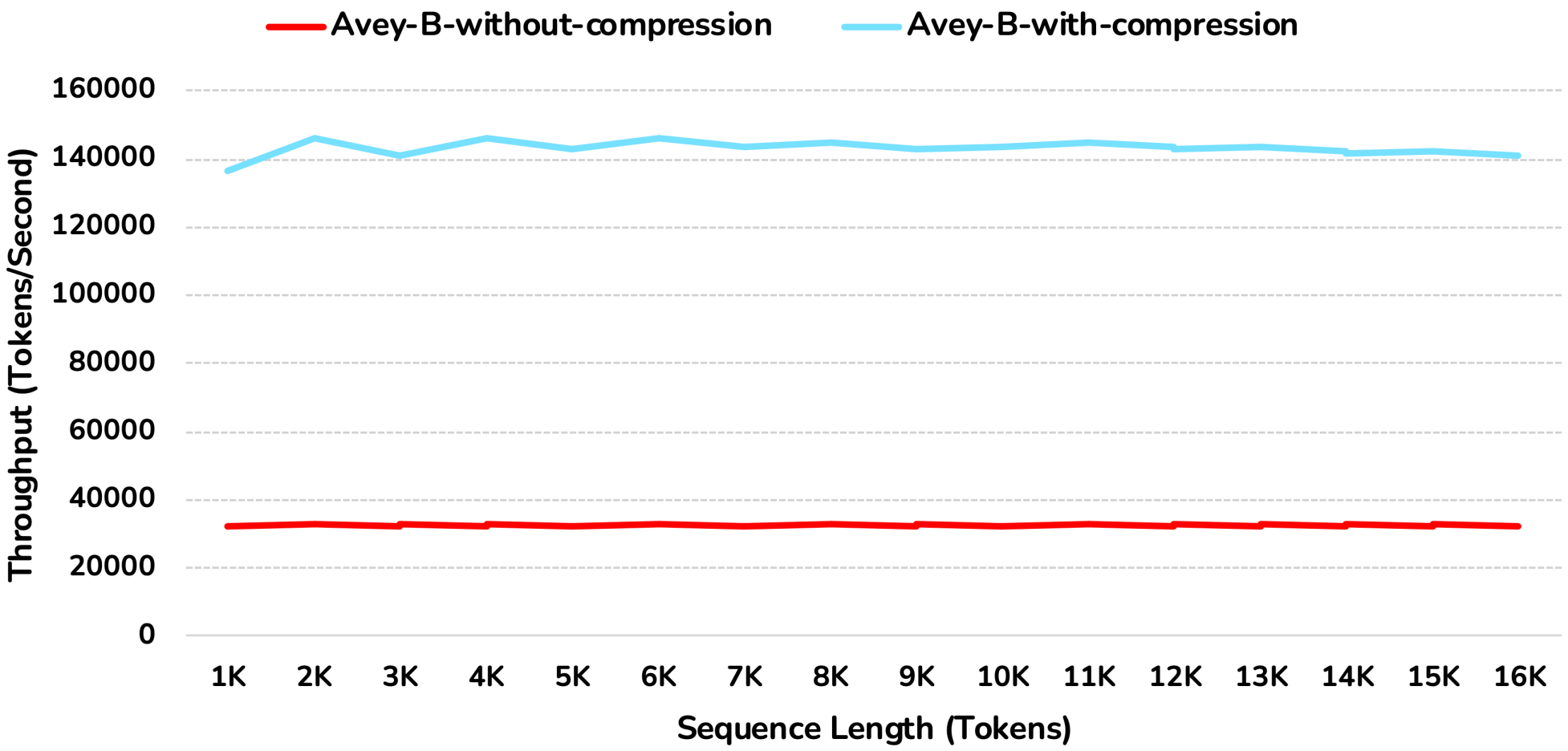
Avey-B: Neural Compression

- **Bidirectional Bottleneck**: The per-split concatenation strategy inflates effective sequence length by a factor of (k+1) in bidirectional regimes
- **Neural Compression**: Avey-B introduces an embedding-wise neural network to condense the concatenated (k+1)S tokens back to S tokens
- **Signal Preservation**: A residual connection is added between the compressor output and the original S tokens to maintain signal integrity



avey

Throughput Improvements with Compression



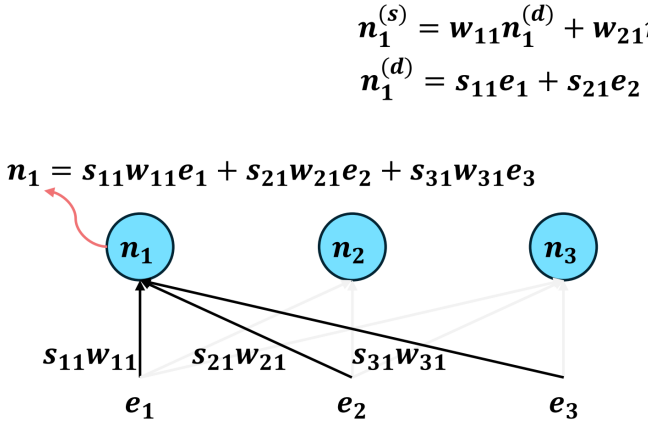
avey

Avey-B: Decoupled Parameterization

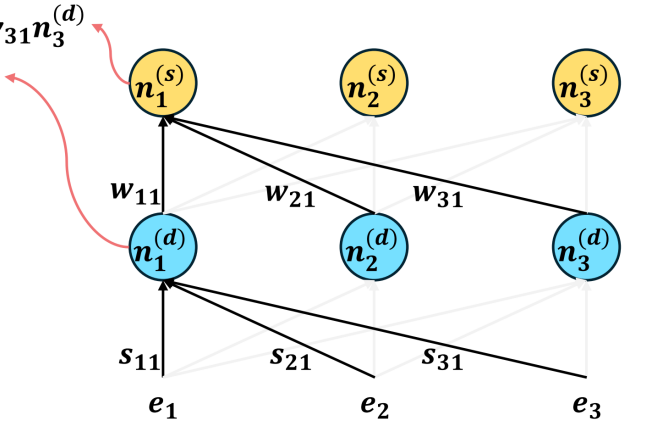
- **The Issue with Coupling**: Coupling static parameters with relevance scores can violate *monotonicity*, causing highly similar tokens to contribute less than less similar ones

- **Interleaved Decoupling**: Avey-B resolves this by separating static layers (learned linear transformations) and dynamic layers (cosine similarity–based weighting) across depth

- **Monotonic Consistency**: This interleaving preserves monotonicity with respect to relevance in the dynamic layers



(a) Avey's *Coupled* Parametrization



(b) Avey-B's *Decoupled* Parametrization

avey

Experimental Setup

Models

Avey-B base
(165M parameters) &
Avey-B large
(391M parameters)

Data

Pre-trained with 180B tokens
from FineWeb

Baselines

BERT, RoBERTa, ModernBERT,
and NeoBERT

Short Range Benchmarks

Sequence Classification
MNLI, QQP, SST-2

Token Classification
CONLL, OntoNotes, UNER

Question Answering
ReCoRD, SQuAD, SQuAD v2

Information Retrieval
MLDR, MS-MARCO, NQ

Long Range Benchmarks

Effectiveness
Synthetic NIAH as extractive QA

Efficiency
Throughput and latency

Short Range Effectiveness

	Model	SC				TC				QA				IR			
		MNLI	QQP	SST-2	**Avg.**	CONLL	Onto.	UNER	**Avg.**	ReCoRD	SQuAD	SQuAD v2	**Avg.**	MLDR	MS MARCO	NQ	**Avg.**
Base	Avey-B	83.58	**89.81**	**92.94**	88.78	**92.88**	**93.80**	**94.10**	**93.59**	44.03	74.44	68.88	62.45	**63.83**	**88.14**	**83.62**	**78.53**
	BERT	81.92	88.57	90.94	87.14	90.25	91.03	88.20	89.82	36.76	72.20	63.99	57.65	57.42	81.15	80.66	73.08
	RoBERTa	86.42	89.12	92.78	89.44	90.55	92.11	88.16	90.27	**67.86**	**80.68**	76.62	**75.05**	56.07	86.47	80.30	74.28
	ModernBERT	**86.72**	89.81	92.32	**89.61**	92.30	93.74	92.30	92.78	65.73	80.23	**77.36**	74.44	54.29	88.09	75.24	72.54
M	NeoBERT	82.53	88.88	84.69	85.36	87.55	88.88	88.17	88.20	37.74	64.84	64.42	55.67	39.98	70.76	59.43	56.72
Large	Avey-B	85.66	89.22	94.38	89.75	**93.60**	**94.09**	**94.32**	**94.00**	58.22	77.30	72.46	69.32	**67.05**	88.72	**86.24**	**80.67**
	BERT	85.08	89.27	92.26	88.87	88.54	90.71	86.09	88.44	52.02	77.93	72.96	67.64	61.08	87.71	85.42	78.07
	RoBERTa	90.16	89.49	94.67	91.44	91.71	92.70	88.79	91.07	**80.86**	**84.00**	**83.04**	**82.63**	58.50	**89.43**	85.91	77.95
	ModernBERT	**90.53**	**90.73**	**95.99**	**92.41**	92.43	93.79	92.92	93.05	73.05	82.02	79.96	78.34	59.64	88.82	81.36	76.61

At the base scale, Avey-B **surpasses BERT and NeoBERT** across all task categories

It also delivers the **strongest results on TC & IR**, outperforming all Transformer-based models

avey

Long Range Effectiveness

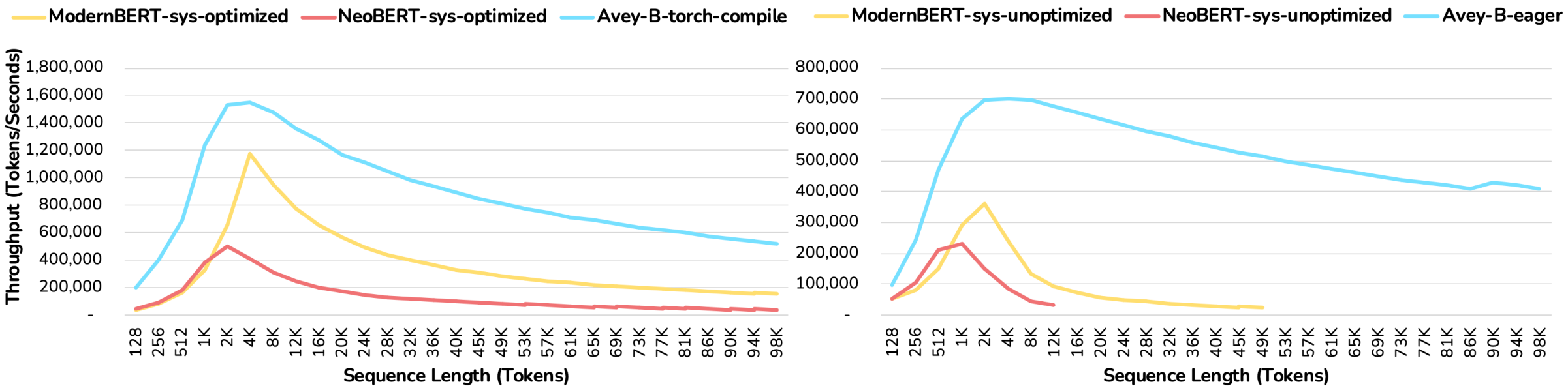
	Model	NIAH-1							
		1k	2k	4k	8k	16k	32k	64k	96k
Base	Avey-B	79.41	79.21	78.94	79.19	78.91	77.73	77.18	75.72
Base	ModernBERT	67.74	67.64	68.31	70.67	–	–	–	–
M	NeoBERT	79.65	79.13	74.73	–	–	–	–	–
Large	Avey-B	79.69	79.24	79.03	79.58	79.44	78.44	76.76	76.06
Large	ModernBERT	68.80	67.52	67.20	OOM	–	–	–	–

	Model	NIAH-2							
		1k	2k	4k	8k	16k	32k	64k	96k
Base	Avey-B	78.29	79.40	79.77	78.53	78.70	75.50	73.78	71.86
Base	ModernBERT	66.99	67.25	69.49	70.48	–	–	–	–
M	NeoBERT	79.61	79.52	80.07	–	–	–	–	–
Large	Avey-B	78.94	79.48	79.99	78.71	79.07	78.31	74.47	74.54
Large	ModernBERT	66.96	67.29	68.07	OOM	–	–	–	–

Avey-B maintains **strong performance well beyond its training sequence length**, and can scale to virtually **unlimited sequence lengths** while remaining efficient

avey

Long Range Efficiency



Both the optimized and unoptimized versions of Avey-B achieve **higher throughput** than their Transformer-based counterparts

This demonstrates both **architectural efficiency** and **practical viability**

avey

Conclusion

Bi-directional Avey

We reformulate the Avey architecture for the bidirectional encoder-only paradigm

Architectural Improvements

We introduce architectural innovations that improve both effectiveness and computational efficiency

SOTA Performance

Avey-B outperforms existing encoder models in both effectiveness and efficiency, despite being trained on substantially fewer tokens

Thank You!

avey