ModernBERT: redefines **encoders**, **SOTA on retrieval and classification**, with support for **long-context** & code, and **record efficiency**

Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference

Benjamin Warner, Antoine Chaffin, Benjamin Clavié
Orion Weller, Oskar Hallström, Said Taghadouini
Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper
Griffin Adams, Jeremy Howard, and Iacopo Poli

Modern Decoder Training Recipe

Trained on 2 Trillion Total Tokens (including web documents, code, and scientific literature)

1.7T Tokens at 1024 context length

Midtraining on 300 Billion Tokens of upsampled higher quality data at 8K context length

Infinite LR schedule for easy continual pretraining

Modern Tokenizer supporting code

Carefully Designed Architecture

Alternating Global-Local Attention to efficiently and accurately process long sequences

Full Model Unpadding and Sequence Packing to not waste compute on padding tokens

Deep and Narrow Design, to balance between downstream performance and hardware efficiency

Hardware-Aware Architecture to maximize throughput on common GPUs

And, last but not least:

Flash Attention & PyTorch**.compile** make the GPUs go brrr

SOTA Across the Board...

Beats DeBERTaV3 on GLUE, first MLM model to ever do so, without any of its tradeoffs

SOTA Retrieval Performance across context lengths, in both single and multi vector settings

Best-In-Class Code Performance thanks to tokenizer and data mixes

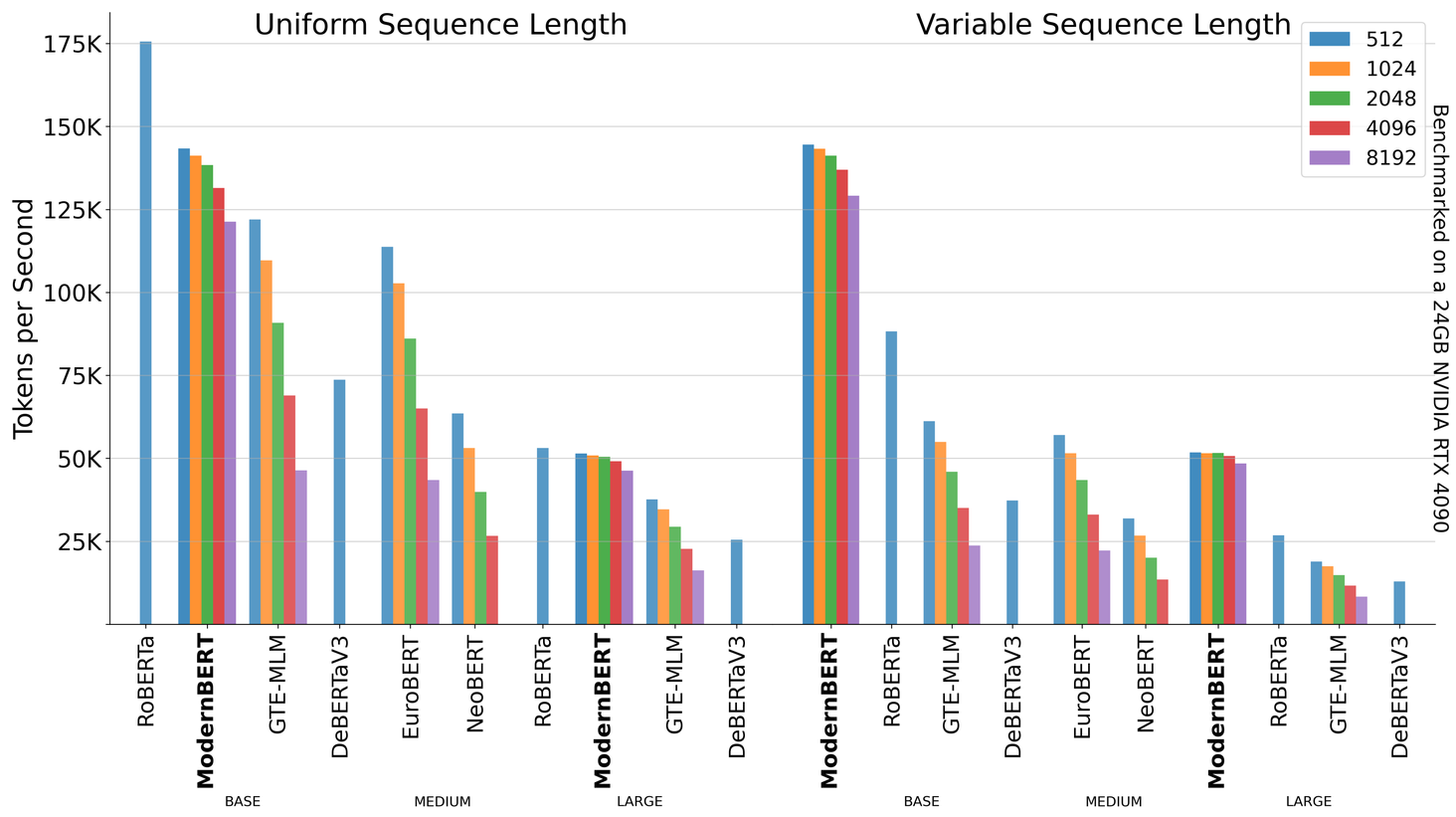
		DPR		ColBERT		NLU	Code	
	Model	BEIR	MLDR	BEIR	MLDR	GLUE	CSN	SQA
Base	BERT	38.9	32.2	49.0	28.1	84.7	41.2	59.5
	RoBERTa	37.7	32.8	48.7	28.2	86.4	44.3	59.6
	DeBERTaV3	20.2	13.4	47.1	21.9	88.1	17.5	18.6
	GTE-en-MLM	41.4	**44.4**	48.2	69.3	85.6	44.9	71.4
	ModernBERT	**41.6**	44.0	**51.3**	**80.2**	**88.4**	**56.4**	**73.6**
Large	BERT	38.9	31.7	49.5	28.5	85.2	41.6	60.8
	RoBERTa	41.4	36.1	49.8	28.8	88.9	47.3	68.1
	DeBERTaV3	25.6	19.2	46.7	23.0	**91.4**	21.2	19.7
	GTE-en-MLM	42.5	**48.9**	50.7	71.3	87.6	40.5	66.9
	ModernBERT	**44.0**	48.6	**52.4**	**80.4**	90.4	**59.5**	**83.9**

...at Light Speed

Blazingly Fast on Short Fixed-Size Inputs, twice as much as DeBERTav3

Variable Sequences Lengths efficient processing thanks to unpadding

Long-Context Class of Its Own: hybrid attention scales to large inputs (over 2× faster than other encoder models)



arXiv

Answer.AI + Lighton