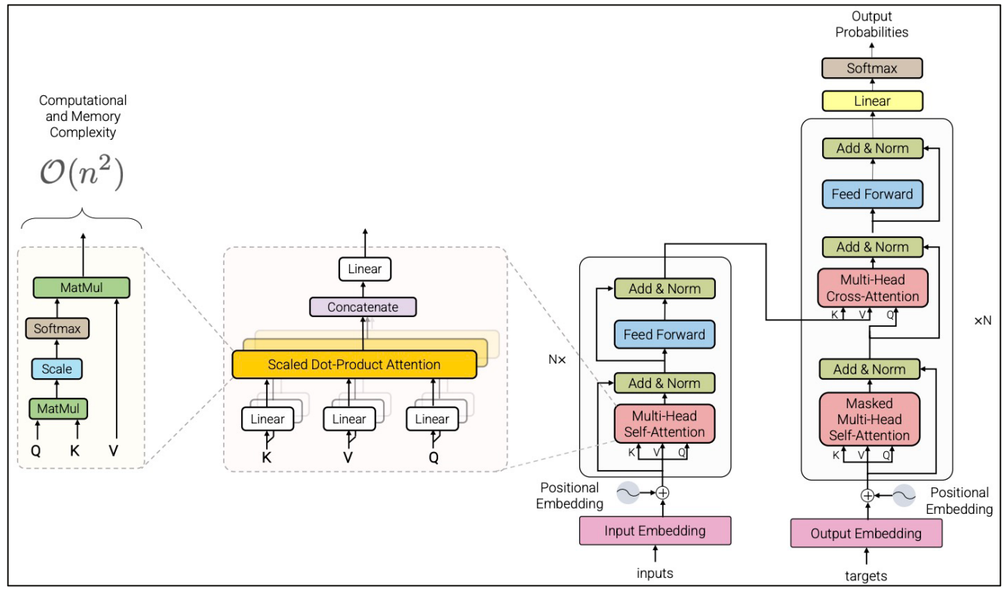
A Quantitative Review on Language Model Efficiency Research (Literature Mining)

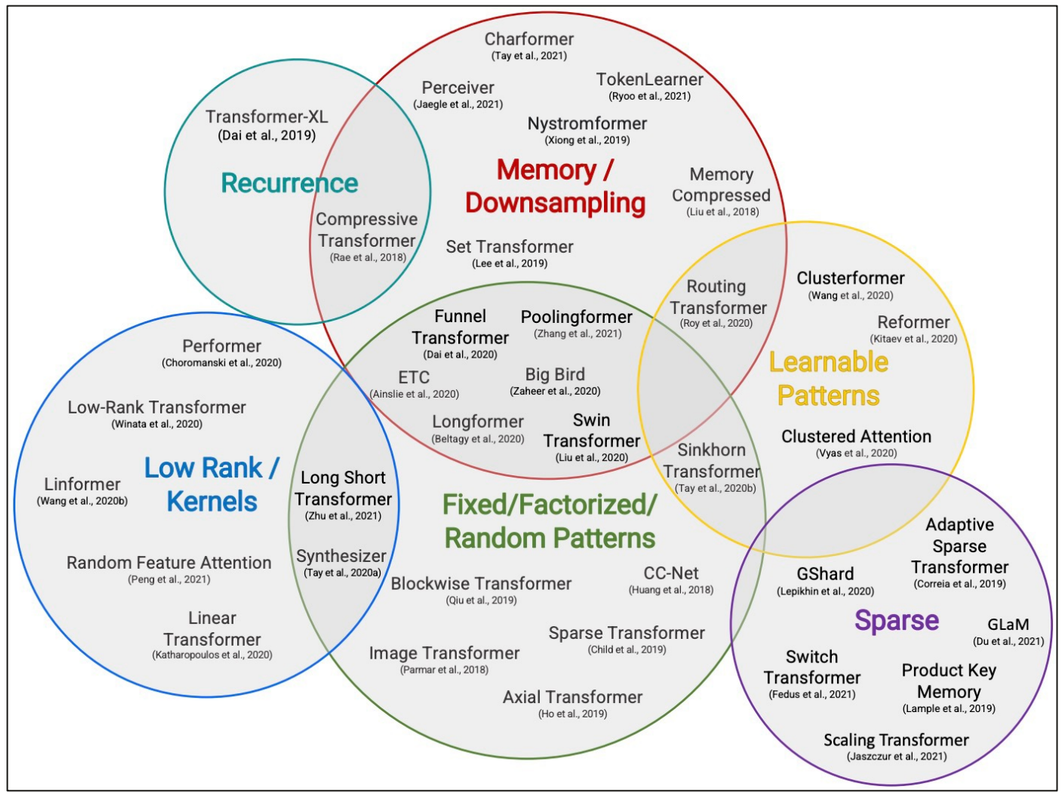
Meng Jiang, Hy Dang, and Lingbo Tong
Department of Computer Science and Engineering
University of Notre Dame

Efficient Transformers: A Survey (2022)

Yi Tay, Mostafa Dehghani, Dara Bahri, Donald Metzler

1. Section 2 *Background*: Introduced standard Transformer architecture.
2. Section 3 *Survey of Models*: Introduced a taxonomy of efficient architectures.
3. Section 4 *Discussion*: “On Evaluation”

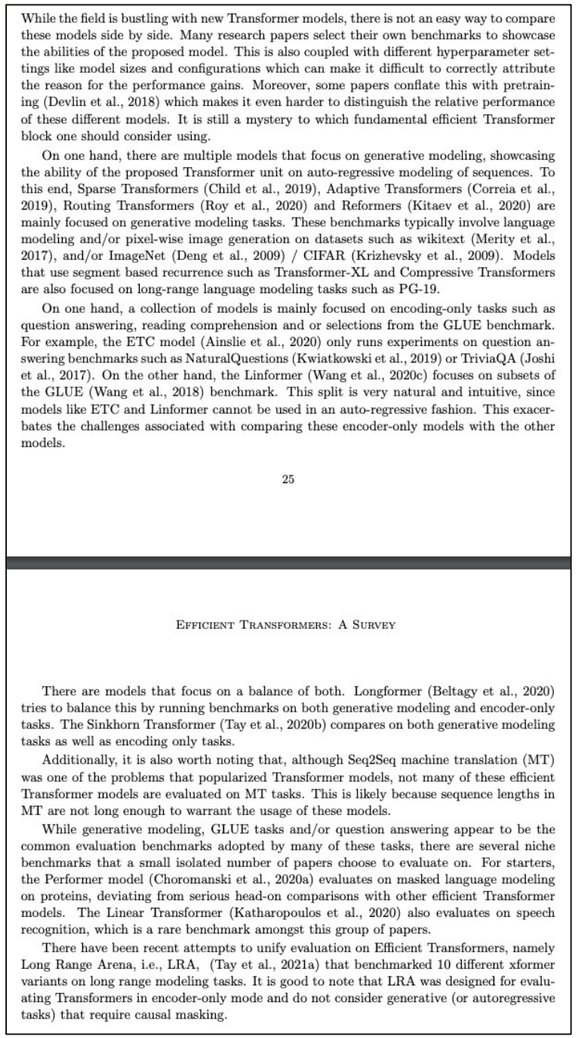




On Evaluation

in Efficient Transformers: A Survey (2022)

1. “Many research papers select their own benchmarks to showcase the abilities of the proposed model.”
2. “This is also coupled with different hyperparameter settings like model sizes and configurations which can make it difficult to correctly attribute the reason for the performance gains.”
3. “Moreover, some papers conflate this with pretraining which makes it even harder to distinguish the relative performance of these different models.”



While the field is bustling with new Transformer models, there is not an easy way to compare these models side by side. Many research papers select their own benchmarks to showcase the abilities of the proposed model. This is also coupled with different hyperparameter settings like model sizes and configurations which can make it difficult to correctly attribute the reason for the performance gains. Moreover, some papers conflate this with pretraining (Devlin et al., 2018) which makes it even harder to distinguish the relative performance of these different models. It is still a mystery to which fundamental efficient Transformer block one should consider using.

On one hand, there are multiple models that focus on generative modeling, showcasing the ability of the proposed Transformer unit on auto-regressive modeling of sequences. To this end, Sparse Transformers (Child et al., 2019), Adaptive Transformers (Correia et al., 2019), Routing Transformers (Roy et al., 2020) and Reformers (Kitaev et al., 2020) are mainly focused on generative modeling tasks. These benchmarks typically involve language modeling and/or pixel-wise image generation on datasets such as wikitext (Merity et al., 2017), and/or ImageNet (Deng et al., 2009) / CIFAR (Krizhevsky et al., 2009). Models that use segment based recurrence such as Transformer-XL and Compressive Transformers are also focused on long-range language modeling tasks such as PG-19.

On one hand, a collection of models is mainly focused on encoding-only tasks such as question answering, reading comprehension and or selections from the GLUE benchmark. For example, the ETC model (Ainslie et al., 2020) only runs experiments on question answering benchmarks such as NaturalQuestions (Kwiatkowski et al., 2019) or TriviaQA (Joshi et al., 2017). On the other hand, the Linformer (Wang et al., 2020c) focuses on subsets of the GLUE (Wang et al., 2018) benchmark. This split is very natural and intuitive, since models like ETC and Linformer cannot be used in an auto-regressive fashion. This exacerbates the challenges associated with comparing these encoder-only models with the other models.

25

Efficient Transformers: A Survey

There are models that focus on a balance of both. Longformer (Beltagy et al., 2020) tries to balance this by running benchmarks on both generative modeling and encoder-only tasks. The Sinkhorn Transformer (Tay et al., 2020b) compares on both generative modeling tasks as well as encoding only tasks.

Additionally, it is also worth noting that, although Seq2Seq machine translation (MT) was one of the problems that popularized Transformer models, not many of these efficient Transformer models are evaluated on MT tasks. This is likely because sequence lengths in MT are not long enough to warrant the usage of these models.

While generative modeling, GLUE tasks and/or question answering appear to be the common evaluation benchmarks adopted by many of these tasks, there are several niche benchmarks that a small isolated number of papers choose to evaluate on. For starters, the Performer model (Choromanski et al., 2020a) evaluates on masked language modeling on proteins, deviating from serious head-on comparisons with other efficient Transformer models. The Linear Transformer (Katharopoulos et al., 2020) also evaluates on speech recognition, which is a rare benchmark amongst this group of papers.

There have been recent attempts to unify evaluation on Efficient Transformers, namely Long Range Arena, i.e., LRA, (Tay et al., 2021a) that benchmarked 10 different xformer variants on long range modeling tasks. It is good to note that LRA was designed for evaluating Transformers in encoder-only mode and do not consider generative (or autoregressive tasks) that require causal masking.

Why Quantitative Review?

on Efficiency Research

1. It's hard to get to know what model was empirically better than the others, but research cannot ignore it: we care what is good, better and the best.
2. Reviewing empirical results from multiple papers (instead of only one) can improve our understanding of the characteristics of models.
3. It can also inspire us to design better research methods.

Key Questions

in this Quantitative Review (and Reading Group Discussion)

1. What are the state-of-the-art efficient LMs?
2. Are the results reported and confirmed by multiple sources? Are there inconsistent results, i.e., significantly different reported performance of one LM reported by different sources?
3. Most studies wanted to claim that theirs were the best solutions when the papers were submitted or accepted. Are these claims correct?

Efficient Language Model Architectures

in this Review

Transformers

- Sparse Transformer [Child et al. 2019]
- Transformer-XL [Dai et al. 2019]
- Reformer [Kitaev et al. 2020]
- Longformer [Beltagy et al. 2020]
- Linear Transformer [Wang et al. 2020]
- SwitchTransformer [Fedus et al. 2022]
- Charformer [Tay et al. 2022]
- ….

State Space Models (SSMs):

- Linear State-Space Layer [Gu et al. 2021]
- Diagonal State Space [Gupta et al. 2022]
- S4: Structured State Space sequence model [Gu et al. 2022]
- S5: Multi-input Multi-output SSMs [Smith et al. 2022]
- ….

Language Tasks, Datasets, Evaluation Metrics

		Bits per character (**BPC**) ↓	**Perplexity** ↓	**Accuracy** ↑	**F1** ↑	**ROUGE-L** ↑	Matthew Correlation Coefficient (**MCC**) ↑	Spearman's rank correlation coefficient (**SC-P/SC-S**) ↑
Language modeling	enwik8	■						
	wikitext		■					
Question answering	TriviaQA			■	■			
	NQ				■			
Summarization	arXiv					■		
GLUE (NLU/NLI)	CoLA						■	
	SST-2			■				
	MRPC			■	■			
	STS-B							■
	QQP			■	■			
	MNLI			■				
	QNLI			■				
	RTE			■				
Long-range arena (LRA)	ListOps			■				
	Text			■				
	Retrieval			■				
	Image			■				
	Pathfinder			■				

Not evaluated on efficiency metrics
but *effectiveness* at a certain level of efficiency

Key Observations

from this Quantitative Review

1. Reducing time and/or memory complexity would inevitably sacrifice a bit non-efficiency performance like accuracy. When the complexities of a set of models were reduced to a certain level, one could claim that the model that achieved the highest accuracy would be the most efficient solution. Past empirical studies unanimously performed efficiency evaluation based on this hypothesis.

Comment: What are the pros and cons of this indirect evaluation method?


Let’s look at results on each task.

Language Modeling — enwik8

Model	BPC ↓	Sources
7L LSTM [Graves, 2013]	1.67	Table 4 in [Rae *et al.*, 2020]
Sliding window	1.34	Table 5 in [Wang *et al.*, 2021]
LN HyperNetworks [Ha *et al.*, 2016]	1.34	Table 2 in [Dai *et al.*, 2019], Table 4 in [Rae *et al.*, 2020]
Locality-Sensitive Hashing [Kitaev *et al.*, 2020]	1.33	Table 5 in [Wang *et al.*, 2021]
LN HM-LSTM [Chung *et al.*, 2016]	1.32	Table 2 in [Dai *et al.*, 2019], Table 4 in [Rae *et al.*, 2020]
ByteNet [Kalchbrenner *et al.*, 2016]	1.31	Table 4 in [Rae *et al.*, 2020]
Sparse Attention [Child *et al.*, 2019]	1.29	Table 5 in [Wang *et al.*, 2021]
RHN [Zilly *et al.*, 2017]	1.27	Table 2 in [Dai *et al.*, 2019], Table 4 in [Rae *et al.*, 2020]
FS-LSTM-4 [Mujika *et al.*, 2017]	1.25	Table 2 in [Dai *et al.*, 2019]
Large mLSTM / mLSTM [Krause *et al.*, 2016]	1.24	Table 2 in [Dai *et al.*, 2019] / Table 4 in [Rae *et al.*, 2020]
cmix v13 [Knol, 2017]	1.23	Table 2 in [Dai *et al.*, 2019]
Cluster-Former (#C=512) [Wang *et al.*, 2021]	**1.22**	Table 5 in [Wang *et al.*, 2021]
T12 [Al-Rfou *et al.*, 2019]	1.11	Table 4 in [Wang *et al.*, 2021]
64L Transformer / 64L Transf. [Al-Rfou *et al.*, 2019]	1.06	Table 2 in [Dai *et al.*, 2019] / Table 4 in [Rae *et al.*, 2020]
Transformer-XL / XFM-XL [Dai *et al.*, 2019]	1.06	Table 5 in [Wang *et al.*, 2021], Table 5 in [Ma *et al.*, 2023]
Reformer [Kitaev *et al.*, 2020]	1.05	Table 4 in [Zhu *et al.*, 2021]
Adaptive [Sukhbaatar *et al.*, 2019]	1.02	Table 5 in [Wang *et al.*, 2021]
MEGA [Ma *et al.*, 2023]	**1.02**	Table 5 in [Ma *et al.*, 2023]
BP-Transformer [Ye *et al.*, 2019]	1.02	Table 5 in [Wang *et al.*, 2021]
Longformer [Tay *et al.*, 2021b]	1.00	Table 5 in [Wang *et al.*, 2021]
24L Transformer-XL / 24L TXL [Dai *et al.*, 2019]	**0.99**	Table 2 in [Dai *et al.*, 2019] / Table 4 in [Rae *et al.*, 2020]
Adaptive Transf. [Sukhbaatar *et al.*, 2019]	0.98	Table 4 in [Rae *et al.*, 2020]
24L Compressive Transformer [Rae *et al.*, 2020]	**0.97**	Table 4 in [Rae *et al.*, 2020]

where the number comes from

multiple: (a) either confirmed or re-used

the same name

different names

"Compressive Transformer" was not in [Want et al. 2021]

multiple: (b) sets of inconsistent reported numbers

Overall state-of-the-art

claimed as "state-of-the-art" (better than any baseline)

Language Modeling – wikitext

Model	PPL ↓	Sources
LSTM	48.7	Table 5 in [Rae *et al.*, 2020], Table 1 in [Dai *et al.*, 2019]
Temporal CNN / TCN [Bai *et al.*, 2018a]	45.2	Table 5 in [Rae *et al.*, 2020] / Table 1 in [Dai *et al.*, 2019]
LSTMs / LSTM + Neural cache [Grave *et al.*, 2016]	40.8	Table 2 in [Roy *et al.*, 2021] / Table 1 in [Dai *et al.*, 2019]
GLU CNN / GCNN-14 / GCNN-14 [Dauphin *et al.*, 2017]	37.2	Table 8 in [Gu *et al.*, 2022b] / Table 5 in [Rae *et al.*, 2020] / Table 1 in [Dai *et al.*, 2019]
AWD-QRNN / Quasi-RNN / QRNNs / QRNN [Bradbury *et al.*, 2016]	33	Table 8 in [Gu *et al.*, 2022b] / Table 5 in [Rae *et al.*, 2020] / Table 2 in [Roy *et al.*, 2021] / Table 1 in [Dai *et al.*, 2019]
RMC [Santoro *et al.*, 2018]	31.9	Table 5 in [Rae *et al.*, 2020]
Hebbian + Cache [Rae *et al.*, 2018]	29.9	Table 1 in [Dai *et al.*, 2019]
LSTM + Hebb. [Rae *et al.*, 2018]	29.2	Table 8 in [Gu *et al.*, 2022b], Table 5 in [Rae *et al.*, 2020]
TrellisNet [Bai *et al.*, 2018b]	29.19	Table 8 in [Gu *et al.*, 2022b]
Transformer [Baevski and Auli, 2018]	26.2	Table 1 in [Peng *et al.*, 2021]
Dynamic Conv [Wu *et al.*, 2019]	25	Table 8 in [Gu *et al.*, 2022b]
RFA [Peng *et al.*, 2021]	**23.5**	Table 1 in [Peng *et al.*, 2021]
TaLK Conv [Lioutas and Guo, 2020]	23.3	Table 8 in [Gu *et al.*, 2022b]
S4 [Gu *et al.*, 2022b]	21.28	Table 8 in [Gu *et al.*, 2022b], Table 5 in [Ma *et al.*, 2023]
Sliding window	20.8	Table 5 in [Wang *et al.*, 2021]
Locality-Sensitive Hashing [Kitaev *et al.*, 2020]	20.8	Table 5 in [Wang *et al.*, 2021]
Adaptive Transformer [Sukhbaatar *et al.*, 2019]	20.6	Table 2 in [Roy *et al.*, 2021]
Transformer [Baevski and Auli, 2018]	20.51	Table 8 in [Gu *et al.*, 2022b]
Sparse Attention [Child *et al.*, 2019]	20.5	Table 5 in [Wang *et al.*, 2021]
Adaptive Input [Baevski and Auli, 2018]	20.5	Table 1 in [Dai *et al.*, 2019]
Clusterformer [Wang *et al.*, 2021]	**20.2**	Table 4 in [Wang *et al.*, 2021]
Local Transformer [Vaswani *et al.*, 2017]	19.8	Table 2 in [Roy *et al.*, 2021]
Transformer / Adaptive Input [Baevski and Auli, 2018]	18.7	Table 5 in [Rae *et al.*, 2020] / Table 2 in [Roy *et al.*, 2021]
XFM-adaptive [Baevski and Auli, 2018; Al-Rfou *et al.*, 2019]	18.66	Table 5 in [Ma *et al.*, 2023]
XFM-XL / 18L TransformerXL / TransformerXL / TransformerXL Large [Dai *et al.*, 2019]	**18.3**	Table 5 in [Ma *et al.*, 2023] / Table 5 in [Rae *et al.*, 2020] / Table 2 in [Roy *et al.*, 2021] / Table 1 in [Dai *et al.*, 2019]
MEGA [Ma *et al.*, 2023]	**18.07**	Table 5 in [Ma *et al.*, 2023]
Compressive Transformer [Rae *et al.*, 2020]	**17.1**	Table 5 in [Rae *et al.*, 2020]
Routing Transformer [Roy *et al.*, 2021]	**15.8**	Table 2 in [Roy *et al.*, 2021]

Question Answering – TriviaQA

includes 950K both human-verified and machine-generated **reading comprehension** question-answer pairs from 662K documents collected from Wikipedia and the web.

Model	Acc	F1	Sources
On **Dev**:			
T5-Base [Raffel *et al.*, 2020]	24.5	-	Table 5 in [Fedus *et al.*, 2022]
T5-Large [Raffel *et al.*, 2020]	29.5	-	Table 5 in [Fedus *et al.*, 2022]
Switch-Base [Fedus *et al.*, 2022]	**30.7**	-	Table 5 in [Fedus *et al.*, 2022]
Switch-Large [Fedus *et al.*, 2022]	**36.9**	-	Table 5 in [Fedus *et al.*, 2022]
Switch-C [Fedus *et al.*, 2022]	**47.5**	-	Table 5 in [Du *et al.*, 2022]
GPT-3 Zero-shot [Brown *et al.*, 2020]	64.3	-	Table 11 in [Du *et al.*, 2022]
GPT-3 One-shot [Brown *et al.*, 2020]	68	-	Table 11 in [Du *et al.*, 2022]
GPT-3 64-shot [Brown *et al.*, 2020]	71.2	-	Table 11 in [Du *et al.*, 2022]
GLaM Zero-shot [Du *et al.*, 2022]	**71.3**	-	Table 11 in [Du *et al.*, 2022]
GLaM One-shot [Du *et al.*, 2022]	**75.8**	-	Table 11 in [Du *et al.*, 2022]
RoBERTa [Liu *et al.*, 2019]	-	74.3	Table 2 in [Zaheer *et al.*, 2020]
Longformer [Beltagy *et al.*, 2020]	-	75.2	Table 2 in [Zaheer *et al.*, 2020]
BIGBIRD-ETC [Zaheer *et al.*, 2020]	-	78.7	Table 2 in [Zaheer *et al.*, 2020]
BIGBIRD-ITC [Zaheer *et al.*, 2020]	-	**79.5**	Table 2 in [Zaheer *et al.*, 2020]
On **Test**:			
KG-FiD (large) [Yu *et al.*, 2021]	69.8	-	Table 5 in [Du *et al.*, 2022]
GPT-3 64-shot [Brown *et al.*, 2020]	71.2	-	Table 5 in [Du *et al.*, 2022]
GLaM One-shot [Du *et al.*, 2022]	75	-	Table 5 in [Du *et al.*, 2022]
RoBERTa-base [Liu *et al.*, 2019]	-	74.3	Table 7 in [Beltagy *et al.*, 2020]
Longformer-base [Beltagy *et al.*, 2020]	-	**75.2**	Table 7 in [Beltagy *et al.*, 2020]
Longformer-large [Beltagy *et al.*, 2020]	-	**77.3**	Table 8 in [Beltagy *et al.*, 2020], Table 3 in [Zaheer *et al.*, 2020]
SpanBERT [Joshi *et al.*, 2020]	-	79.1	Table 3 in [Zaheer *et al.*, 2020]
Fusion-in-Decoder [Izacard and Grave, 2020]	-	84.4	Table 3 in [Zaheer *et al.*, 2020]
BIGBIRD-ETC [Zaheer *et al.*, 2020]	-	**84.5**	Table 3 in [Zaheer *et al.*, 2020]
On **Test Verified**:			
Longformer [Beltagy *et al.*, 2020]	-	85.3	Table 3 in [Zaheer *et al.*, 2020]
SpanBERT [Joshi *et al.*, 2020]	-	86.6	Table 3 in [Zaheer *et al.*, 2020]
Fusion-in-Decoder [Izacard and Grave, 2020]	-	90.3	Table 3 in [Zaheer *et al.*, 2020]
BIGBIRD-ETC [Zaheer *et al.*, 2020]	-	**92.4**	Table 3 in [Zaheer *et al.*, 2020]

Question Answering – NQ

contains 307,373 training examples, 7,830 development examples, and 7,842 test examples.
Each example is comprised of a **google.com query** and a Wikipedia page.

Model	F1 ↑	Sources
On **Dev**:		
DecAtt [Parikh *et al.*, 2016] + DocReader [Chen *et al.*, 2017]	54.8	Table 2 in [Zhang *et al.*, 2021], Table 2 in [Wang *et al.*, 2021]
BERT-large / BERT-large / BERT-joint [Alberti *et al.*, 2019]	64.7	Table 2 in [Ainslie *et al.*, 2020] / Table 2 in [Zhang *et al.*, 2021] / Table 2 in [Wang *et al.*, 2021]
Sparse Transformer / Sparse Attention [Jaszczur *et al.*, 2021]	74.5	Table 2 in [Zhang *et al.*, 2021] / Table 2 in [Wang *et al.*, 2021]
RikiNet-RoBERTa / RikiNet / RikiNet [Liu *et al.*, 2020]	75.3	Table 2 in [Wang *et al.*, 2021] / Table 2 in [Ainslie *et al.*, 2020] / Table 2 in [Zhang *et al.*, 2021]
Reformer / Locality-Sensitive Hashing [Kitaev *et al.*, 2020]	75.5	Table 2 in [Zhang *et al.*, 2021] / Table 2 in [Wang *et al.*, 2021]
RikiNet-ensemble [Liu *et al.*, 2020]	**75.9**	Table 2 in [Zhang *et al.*, 2021]
Cluster-Former [Wang *et al.*, 2021]	**76.5**	Table 2 in [Zhang *et al.*, 2021], Table 2 in [Wang *et al.*, 2021]
ReflectionNet-ensemble [Wang *et al.*, 2020b]	**77.0**	Table 2 in [Zhang *et al.*, 2021]
Poolingformer [Zhang *et al.*, 2021]	**77.5**	Table 2 in [Zhang *et al.*, 2021]
ETC-large (lifting from RoBERTa) [Ainslie *et al.*, 2020]	**78.2**	Table 2 in [Ainslie *et al.*, 2020]
On **Test**:		
RikiNet-v2 / RikiNet-ensemble [Liu *et al.*, 2020]	76.1	Table 3 in [Zaheer *et al.*, 2020] / Table 2 in [Zhang *et al.*, 2021]
ReflectionNet [Liu *et al.*, 2020]	77.1	Table 3 in [Zaheer *et al.*, 2020]
ReflectionNet-ensemble [Liu *et al.*, 2020]	77.2	Table 2 in [Zhang *et al.*, 2021]
ETC (official) [Ainslie *et al.*, 2020]	**77.78**	Table 5 in [Ainslie *et al.*, 2020]
BIGBIRD-ETC [Zaheer *et al.*, 2020]	**77.8**	Table 3 in [Zaheer *et al.*, 2020], Table 2 in [Zhang *et al.*, 2021]
Cluster-Former-ensemble / Cluster-Former [Wang *et al.*, 2021]	**78**	Table 3 in [Wang *et al.*, 2021] / Table 2 in [Zhang *et al.*, 2021]
Poolingformer-ensamble [Zhang *et al.*, 2021]	**79.8**	Table 2 in [Zhang *et al.*, 2021]

Summarization – arXiv

Model	R-L ↑	Sources
Long-Doc-Seq2Seq / Discourse-aware [Cohan *et al.*, 2018]	31.80	Table 4 in [Zaheer *et al.*, 2020] / Table 11 in [Beltagy *et al.*, 2020]
Extr-Abst-TLM [Subramanian *et al.*, 2019]	38.03	Table 4 in [Zaheer *et al.*, 2020], Table 11 in [Beltagy *et al.*, 2020], Table 4 in [Zhang *et al.*, 2021]
Sent-PTR[Subramanian *et al.*, 2019]	38.06	Table 4 in [Zaheer *et al.*, 2020], Table 4 in [Zhang *et al.*, 2021]
Dancer / Dancer / Dancer RUM [Gidiotis and Tsoumakas, 2020]	38.44	Table 4 in [Zaheer *et al.*, 2020] / Table 4 in [Zhang *et al.*, 2021] / Table 5 in [Jaszczur *et al.*, 2021]
Pegasus [Zhang *et al.*, 2020]	38.83	Table 4 in [Zaheer *et al.*, 2020], Table 5 in [Jaszczur *et al.*, 2021], Table 11 in [Beltagy *et al.*, 2020], Table 4 in [Zhang *et al.*, 2021]
Pegasus (Re Eval) [Zhang *et al.*, 2020]	39.17	Table 4 in [Zaheer *et al.*, 2020]
Dancer / Dancer PEGASUS [Gidiotis and Tsoumakas, 2020]	40.56	Table 4 in [Zhang *et al.*, 2021] / Table 5 in [Jaszczur *et al.*, 2021]
BIGBIRD-Pegasus / BIGBIRD-Pegasus / BigBird (seqlen:4096) / BigBird [Zaheer *et al.*, 2020]	**41.77**	Table 4 in [Zaheer *et al.*, 2020] / Table 5 in [Jaszczur *et al.*, 2021] / Table 11 in [Beltagy *et al.*, 2020] / Table 4 in [Zhang *et al.*, 2021]
LED-large (seqlen: 16384) / LED16k [Beltagy *et al.*, 2020]	**41.83**	Table 11 in [Beltagy *et al.*, 2020] / Table 4 in [Zhang *et al.*, 2021]
Poolingformer16k [Zhang *et al.*, 2021]	**42.69**	Table 4 in [Zhang *et al.*, 2021]

GLUE – MNLI (Multi-Genre Natural Language Inference)

Model	Acc (m) ↑	Acc (mm)	Sources
DyConv [Wu *et al.*, 2019]	73.8	75.1	Table 5 in [Tay *et al.*, 2020a]
Nystromformer [Xiong *et al.*, 2021]	80.9	82.2	Table 2 in [Xiong *et al.*, 2021]
BERT-base[Devlin *et al.*, 2018]	82.4	82.4	Table 2 in [Xiong *et al.*, 2021]
BERT [Devlin *et al.*, 2018]	84.6	83.4	Table 16 in [Zaheer *et al.*, 2020]
T5-Base [Raffel *et al.*, 2020]	84.7	85	Table 5 in [Tay *et al.*, 2020a]
Syn (R+V) [Tay *et al.*, 2020a]	**85**	**84.6**	Table 5 in [Tay *et al.*, 2020a]
XLNet [Yang *et al.*, 2019]	86.8	-	Table 16 in [Zaheer *et al.*, 2020]
BIGBIRD [Zaheer *et al.*, 2020]	87.5	87.3	Table 16 in [Zaheer *et al.*, 2020]
ROBERTABase / RoBERTa [Liu *et al.*, 2019]	87.6	-	Table 3 in [Dai *et al.*, 2020] / Table 16 in [Zaheer *et al.*, 2020]
MPNetBase [Song *et al.*, 2020]	88.5	-	Table 3 in [Dai *et al.*, 2020]
Transformer [Vaswani *et al.*, 2017]	89.4	-	Table 2 in [Dai *et al.*, 2020]
ROBERTALarge [Liu *et al.*, 2019]	90.2	-	Table 3 in [Dai *et al.*, 2020]
XLNetLarge [Yang *et al.*, 2019]	90.8	-	Table 3 in [Dai *et al.*, 2020]
ELECTRALarge [Clark *et al.*, 2020]	90.9	-	Table 3 in [Dai *et al.*, 2020]
Funnel Transformer (B10-10-10H1024) [Dai *et al.*, 2020]	**91.1**	-	Table 3 in [Dai *et al.*, 2020]

GLUE – SST-2 (Stanford Sentiment Treebank)

Model	Acc ↑	Sources
BERT-base [Devlin *et al.*, 2018]	90	Table 2 in [Xiong *et al.*, 2021]
DyConv [Wu *et al.*, 2019]	90.6	Table 5 in [Tay *et al.*, 2020a]
Nystromformer [Xiong *et al.*, 2021]	**91.4**	Table 2 in [Xiong *et al.*, 2021]
Syn (D+V) [Tay *et al.*, 2020a]	**92.4**	Table 5 in [Tay *et al.*, 2020a]
Linformer-128 (16GB) [Wang *et al.*, 2020a]	92.4	Table 6 in [Ma *et al.*, 2021]
BERT-base [Devlin *et al.*, 2018]	92.7	Table 6 in [Ma *et al.*, 2021]
T5-Base+[Raffel *et al.*, 2020]	92.9	Table 5 in [Tay *et al.*, 2020a]
BERT [Devlin *et al.*, 2018]	93.5	Table 16 in [Zaheer *et al.*, 2020]
BIGBIRD [Zaheer *et al.*, 2020]	94.6	Table 16 in [Zaheer *et al.*, 2020]
Luna-128 (160GB) [Ma *et al.*, 2021]	94.6	Table 6 in [Ma *et al.*, 2021]
XLNet [Yang *et al.*, 2019]	94.7	Table 16 in [Zaheer *et al.*, 2020]
Transformer (L24H1024) [Vaswani *et al.*, 2017]	94.8	Table 2 in [Dai *et al.*, 2020]
ROBERTABase / RoBERTa / RoBERTa-base (160GB) [Liu *et al.*, 2019]	94.8	Table 3 in [Dai *et al.*, 2020] / Table 16 in [Zaheer *et al.*, 2020] / Table 6 in [Ma *et al.*, 2021]
MPNetBase [Song *et al.*, 2020]	95.4	Table 3 in [Dai *et al.*, 2020]
ROBERTALarge [Liu *et al.*, 2019]	96.4	Table 3 in [Dai *et al.*, 2020]
Funnel Transformer (B10-10-10H1024) [Dai *et al.*, 2020]	96.8	Table 3 in [Dai *et al.*, 2020]
ELECTRALarge [Clark *et al.*, 2020]	96.9	Table 3 in [Dai *et al.*, 2020]
XLNetLarge [Yang *et al.*, 2019]	97	Table 3 in [Dai *et al.*, 2020]

GLUE – QNLI (Question-answering Natural Language Inference)

Model	Acc ↑	Sources
DyConv [Wu *et al.*, 2019]	84.4	Table 5 in [Tay *et al.*, 2020a]
BERT-base [Devlin *et al.*, 2018]	88.4	Table 6 in [Ma *et al.*, 2021]
Nystromformer [Xiong *et al.*, 2021]	88.7	Table 2 in [Xiong *et al.*, 2021]
BERT-base [Devlin *et al.*, 2018]	90.3	Table 2 in [Xiong *et al.*, 2021]
Linformer-128 (16GB) [Wang *et al.*, 2020a]	90.4	Table 6 in [Ma *et al.*, 2021]
BERT [Devlin *et al.*, 2018]	90.5	Table 16 in [Zaheer *et al.*, 2020]
T5-Base [Raffel *et al.*, 2020]	91.7	Table 5 in [Tay *et al.*, 2020a]
XLNet [Yang *et al.*, 2019]	91.7	Table 16 in [Zaheer *et al.*, 2020]
BIGBIRD [Zaheer *et al.*, 2020]	92.2	Table 16 in [Zaheer *et al.*, 2020]
Luna-128 (160GB) [Ma *et al.*, 2021]	92.2	Table 6 in [Ma *et al.*, 2021]
Syn (R+V) [Tay *et al.*, 2020a]	**92.3**	Table 5 in [Tay *et al.*, 2020a]
ROBERTABase / RoBERTa / RoBERTa-base [Liu *et al.*, 2019]	92.8	Table 3 in [Dai *et al.*, 2020] / Table 16 in [Zaheer *et al.*, 2020] / Table 6 in [Ma *et al.*, 2021]
MPNetBase [Song *et al.*, 2020]	93.3	Table 3 in [Dai *et al.*, 2020]
Transformer (L24H1024)[Vaswani *et al.*, 2017]	94.1	Table 2 in [Dai *et al.*, 2020]
ROBERTALarge [Liu *et al.*, 2019]	94.7	Table 3 in [Dai *et al.*, 2020]
XLNetLarge [Yang *et al.*, 2019]	94.9	Table 3 in [Dai *et al.*, 2020]
ELECTRALarge [Clark *et al.*, 2020]	95	Table 3 in [Dai *et al.*, 2020]
Funnel Transformer (B10-10-10H1024) [Dai *et al.*, 2020]	**95.1**	Table 3 in [Dai *et al.*, 2020]

GLUE – RTE (Recognizing Textual Entailment)

Model	Acc ↑	Sources
DyConv [Wu *et al.*, 2019]	58.1	Table 5 in [Tay *et al.*, 2020a]
BERT [Devlin *et al.*, 2018]	66.4	Table 16 in [Zaheer *et al.*, 2020]
XLNet [Yang *et al.*, 2019]	74	Table 16 in [Zaheer *et al.*, 2020]
BIGBIRD [Zaheer *et al.*, 2020]	75	Table 16 in [Zaheer *et al.*, 2020]
ROBERTABase / RoBERTa [Liu *et al.*, 2019]	78.7	Table 3 in [Dai *et al.*, 2020] / Table 16 in [Zaheer *et al.*, 2020]
T5-Base+ [Raffel *et al.*, 2020]	79.1	Table 5 in [Tay *et al.*, 2020a]
Syn (R+V) [Tay *et al.*, 2020a]	**81.2**	Table 5 in [Tay *et al.*, 2020a]
Transformer (L24H1024) [Vaswani *et al.*, 2017]	84.5	Table 2 in [Dai *et al.*, 2020]
MPNetBase [Song *et al.*, 2020]	85.2	Table 3 in [Dai *et al.*, 2020]
XLNetLarge [Yang *et al.*, 2019]	85.9	Table 3 in [Dai *et al.*, 2020]
ROBERTALarge [Liu *et al.*, 2019]	86.6	Table 3 in [Dai *et al.*, 2020]
ELECTRALarge [Clark *et al.*, 2020]	88	Table 3 in [Dai *et al.*, 2020]
Funnel Transformer (B10-10-10H1024) [Dai *et al.*, 2020]	**89.5**	Table 3 in [Dai *et al.*, 2020]

GLUE – COLA (Corpus of Linguistic Acceptability)

MCC: Matthew Correlation Coefficient

Model	MCC ↑	Sources
DyConv [Wu *et al.*, 2019]	33.9	Table 5 in [Tay *et al.*, 2020a]
BERT [Devlin *et al.*, 2018]	52.1	Table 16 in [Zaheer *et al.*, 2020]
Syn (R+V) [Tay *et al.*, 2020a]	**53.3**	Table 5 in [Tay *et al.*, 2020a]
T5-Base+ [Raffel *et al.*, 2020]	54.3	Table 5 in [Tay *et al.*, 2020a]
BIGBIRD [Zaheer *et al.*, 2020]	58.5	Table 16 in [Zaheer *et al.*, 2020]
XLNet [Yang *et al.*, 2019]	60.2	Table 16 in [Zaheer *et al.*, 2020]
ROBERTABase / RoBERTa [Liu *et al.*, 2019]	63.6	Table 3 in [Dai *et al.*, 2020] / Table 16 in [Zaheer *et al.*, 2020]
MPNetBase [Song *et al.*, 2020]	65	Table 3 in [Dai *et al.*, 2020]
Transformer (L24H1024) [Vaswani *et al.*, 2017]	66.5	Table 2 in [Dai *et al.*, 2020]
ROBERTALarge [Liu *et al.*, 2019]	68	Table 3 in [Dai *et al.*, 2020]
XLNetLarge [Yang *et al.*, 2019]	69	Table 3 in [Dai *et al.*, 2020]
ELECTRALarge [Clark *et al.*, 2020]	69.1	Table 3 in [Dai *et al.*, 2020]
Funnel Transformer (B10-10-10H1024) [Dai *et al.*, 2020]	**88.7**	Table 3 in [Dai *et al.*, 2020]

GLUE – SST-B (Semantic Textual Similarity Benchmark)

SC: Spearman's Correlation Coefficient

Model	SC-P ↑	SC-S	Sources
DyConv [Wu *et al.*, 2019]	60.7	63.1	Table 5 in [Tay *et al.*, 2020a]
BERT [Devlin *et al.*, 2018]	85.8	-	Table 16 in [Zaheer *et al.*, 2020]
BIGBIRD [Zaheer *et al.*, 2020]	87.8	-	Table 16 in [Zaheer *et al.*, 2020]
T5-Base [Raffel *et al.*, 2020]	89.1	88.9	Table 5 in [Tay *et al.*, 2020a]
Syn (R+V) [Tay *et al.*, 2020a]	**89.3**	**88.9**	Table 5 in [Tay *et al.*, 2020a]
XLNet [Yang *et al.*, 2019]	89.5	-	Table 16 in [Zaheer *et al.*, 2020]
MPNetBase [Song *et al.*, 2020]	90.9	-	Table 3 in [Dai *et al.*, 2020]
ROBERTABase / RoBERTa [Liu *et al.*, 2019]	91.2	-	Table 3 in [Dai *et al.*, 2020] / Table 16 in [Zaheer *et al.*, 2020]
Transformer (L24H1024) [Vaswani *et al.*, 2017]	91.5	-	Table 2 in [Dai *et al.*, 2020]
Funnel Transformer (B10-10-10H1024) [Dai *et al.*, 2020]	92.1	-	Table 3 in [Dai *et al.*, 2020]
ROBERTALarge [Liu *et al.*, 2019]	92.4	-	Table 3 in [Dai *et al.*, 2020]
XLNetLarge [Yang *et al.*, 2019]	92.5	-	Table 3 in [Dai *et al.*, 2020]
ELECTRALarge [Clark *et al.*, 2020]	92.6	-	Table 3 in [Dai *et al.*, 2020]

LRA – ListOps

BigBird [Zaheer *et al.*, 2020]	36.05	Table 1 in [Ma *et al.*, 2021], Table 2 in [Ma *et al.*, 2023], Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
XFM / Transformer / Transformer / Transformer / Transformer [Vaswani *et al.*, 2017]	36.37	Table 2 in [Ma *et al.*, 2023] / Table 1 in [Hasani *et al.*, 2022] / Table 1 in [Ma *et al.*, 2021] / Table 10 in [Gu *et al.*, 2022b] / Table 2 in [Smith *et al.*, 2022]
Reformer [Kitaev *et al.*, 2020]	36.44	Table 1 in [Zhu *et al.*, 2021]
Synthesizer [Tay *et al.*, 2021a]	36.99	Table 1 in [Ma *et al.*, 2021], Table 10 in [Gu *et al.*, 2022b]
Standard [Vaswani *et al.*, 2017]	37.1	Table 3 in [Xiong *et al.*, 2021]
Transformer (re-impl) / XFM (re-impl) [Vaswani *et al.*, 2017]	37.11	Table 1 in [Ma *et al.*, 2021] / Table 2 in [Ma *et al.*, 2023]
Full Attention [Vaswani *et al.*, 2017]	37.13	Table 1 in [Zhu *et al.*, 2021]
Nystromformer [Xiong *et al.*, 2021]	37.15	Table 3 in [Xiong *et al.*, 2021], Table 1 in [Hasani *et al.*, 2022], Table 10 in [Gu *et al.*, 2022b]
Linformer [Wang *et al.*, 2020a]	37.25	Table 3 in [Xiong *et al.*, 2021]
Luna-256 [Ma *et al.*, 2021]	37.25	Table 10 in [Gu *et al.*, 2022b], Table 2 in [Ma *et al.*, 2023], Table 1 in [Hasani *et al.*, 2022]
Reformer [Kitaev *et al.*, 2020]	37.27	Table 1 in [Hasani *et al.*, 2022], Table 10 in [Gu *et al.*, 2022b], Table 2 in [Ma *et al.*, 2023], Table 1 in [Ma *et al.*, 2021]
Nystromformer [Xiong *et al.*, 2021]	37.34	Table 1 in [Zhu *et al.*, 2021]
Linformer [Wang *et al.*, 2020a]	37.38	Table 1 in [Zhu *et al.*, 2021]
Luna-256 [Ma *et al.*, 2021]	37.98	Table 1 in [Ma *et al.*, 2021], Table 2 in [Ma *et al.*, 2023]
Luna-128 [Ma *et al.*, 2021]	**38.01**	Table 1 in [Ma *et al.*, 2021]
Transformer-LS [Zhu *et al.*, 2021]	**38.36**	Table 1 in [Zhu *et al.*, 2021]
CCNN [Romero *et al.*, 2022]	43.6	Table 2 in [Smith *et al.*, 2022]
CDIL [Cheng *et al.*, 2023]	44.05	Table 1 in [Hasani *et al.*, 2022]
H-Trans.-1D / H-Transformer-1D [Zhu and Soricut, 2021]	49.53	Table 2 in [Smith *et al.*, 2022] / Table 1 in [Hasani *et al.*, 2022]
DSS [Gupta *et al.*, 2022]	57.6	Table 1 in [Hasani *et al.*, 2022]
S4-v2 (re-impl) [Gu *et al.*, 2022b]	59.1	Table 2 in [Ma *et al.*, 2023]
S4-v2 / S4 (updated) [Gu *et al.*, 2022b]	**59.6**	Table 2 in [Ma *et al.*, 2023] / Table 10 in [Gu *et al.*, 2022b]
S4D-LegS [Gu *et al.*, 2022a]	60.47	Table 2 in [Smith *et al.*, 2022], Table 1 in [Hasani *et al.*, 2022]
S4D-Lin [Gu *et al.*, 2022a]	60.52	Table 1 in [Hasani *et al.*, 2022]
S5 [Smith *et al.*, 2022]	62.15	Table 2 in [Smith *et al.*, 2022], Table 1 in [Hasani *et al.*, 2022]
Liquid-S4 / Liquid-S4-PB [Hasani *et al.*, 2022]	**62.75**	Table 2 in [Smith *et al.*, 2022] / Table 1 in [Hasani *et al.*, 2022]
MEGA [Ma *et al.*, 2023]	**63.14**	Table 2 in [Ma *et al.*, 2023], Table 2 in [Smith *et al.*, 2022]

LRA – ListOps *(cont.)*

Model	Acc ↑	Sources
Local Attention / Local Attention / Local Attn. [Tay *et al.*, 2020c]	15.82	Table 1 in [Ma *et al.*, 2021] / Table 10 in [Gu *et al.*, 2022b] / Table 1 in [Hasani *et al.*, 2022]
Linear Trans. [Katharopoulos *et al.*, 2020]	16.13	Table 1 in [Ma *et al.*, 2021], Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
Linformer [Wang *et al.*, 2020a]	16.13	Table 1 in [Hasani *et al.*, 2022]
Sparse Trans. / Sparse Trans. / Sparse Transformer [Child *et al.*, 2019]	17.07	Table 1 in [Ma *et al.*, 2021] / Table 10 in [Gu *et al.*, 2022b] / Table 1 in [Hasani *et al.*, 2022]
Performer in [Choromanski *et al.*, 2020]	18.01	Table 1 in [Ma *et al.*, 2021], Table 2 in [Ma *et al.*, 2023], Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022], Table 3 in [Xiong *et al.*, 2021]
Reformer [Kitaev *et al.*, 2020]	19.05	Table 3 in [Xiong *et al.*, 2021]
Performer in [Choromanski *et al.*, 2020]	32.78	Table 1 in [Zhu *et al.*, 2021]
Sinkhorn Trans. [Tay *et al.*, 2020b]	33.67	Table 1 in [Ma *et al.*, 2021], Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
FNet [Lee-Thorp *et al.*, 2021]	35.33	Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
Longformer [Beltagy *et al.*, 2020]	35.63	Table 1 in [Ma *et al.*, 2021], Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
Linformer [Wang *et al.*, 2020a]	35.7	Table 1 in [Ma *et al.*, 2021], Table 2 in [Ma *et al.*, 2023], Table 10 in [Gu *et al.*, 2022b]
BigBird [Zaheer *et al.*, 2020]	36.05	Table 1 in [Ma *et al.*, 2021], Table 2 in [Ma *et al.*, 2023], Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
XFM / Transformer / Transformer / Transformer / Transformer [Vaswani *et al.*, 2017]	36.37	Table 2 in [Ma *et al.*, 2023] / Table 1 in [Hasani *et al.*, 2022] / Table 1 in [Ma *et al.*, 2021] / Table 10 in [Gu *et al.*, 2022b] / Table 2 in [Smith *et al.*, 2022]
Reformer [Kitaev *et al.*, 2020]	36.44	Table 1 in [Zhu *et al.*, 2021]
Synthesizer [Tay *et al.*, 2021a]	36.99	Table 1 in [Ma *et al.*, 2021], Table 10 in [Gu *et al.*, 2022b]
Standard [Vaswani *et al.*, 2017]	37.1	Table 3 in [Xiong *et al.*, 2021]
Transformer (re-impl) / XFM (re-impl) [Vaswani *et al.*, 2017]	37.11	Table 1 in [Ma *et al.*, 2021] / Table 2 in [Ma *et al.*, 2023]
Full Attention [Vaswani *et al.*, 2017]	37.13	Table 1 in [Zhu *et al.*, 2021]
Nystromformer [Xiong *et al.*, 2021]	37.15	Table 3 in [Xiong *et al.*, 2021], Table 1 in [Hasani *et al.*, 2022], Table 10 in [Gu *et al.*, 2022b]

LRA – Text

Luna-256 [Ma *et al.*, 2021]	64.57	Table 10 in [Gu *et al.*, 2022b], Table 2 in [Smith *et al.*, 2022], Table 1 in [Hasani *et al.*, 2022]
Reformer [Kitaev *et al.*, 2020]	64.88	Table 1 in [Zhu *et al.*, 2021], Table 3 in [Xiong *et al.*, 2021]
Standard [Vaswani *et al.*, 2017]	65.02	Table 3 in [Xiong *et al.*, 2021]
FNet [Lee-Thorp *et al.*, 2021]	65.11	Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
Performer in [Choromanski *et al.*, 2020]	65.21	Table 1 in [Zhu *et al.*, 2021]
Transformer (re-impl) / XFM (re-impl) [Vaswani *et al.*, 2017]	65.21	Table 1 in [Ma *et al.*, 2021] / Table 2 in [Ma *et al.*, 2023]
Full Attention [Vaswani *et al.*, 2017]	65.35	Table 1 in [Zhu *et al.*, 2021]
Performer in [Choromanski *et al.*, 2020]	65.4	Table 1 in [Ma *et al.*, 2021], Table 2 in [Ma *et al.*, 2023], Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
Nystromformer [Xiong *et al.*, 2021]	**65.52**	Table 3 in [Xiong *et al.*, 2021], Table 1 in [Hasani *et al.*, 2022], Table 10 in [Gu *et al.*, 2022b]
Nystromformer [Xiong *et al.*, 2021]	65.75	Table 1 in [Zhu *et al.*, 2021]
Luna-256 [Ma *et al.*, 2021]	**65.78**	Table 1 in [Ma *et al.*, 2021], Table 2 in [Ma *et al.*, 2023]
Linear Trans. [Katharopoulos *et al.*, 2020]	65.9	Table 1 in [Ma *et al.*, 2021], Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
Linformer [Wang *et al.*, 2020a]	65.9	Table 1 in [Hasani *et al.*, 2022]
Transformer-LS [Zhu *et al.*, 2021]	**68.4**	Table 1 in [Zhu *et al.*, 2021]
DSS [Gupta *et al.*, 2022]	76.6	Table 1 in [Hasani *et al.*, 2022]
H-Trans.-1D / H-Transformer-1D [Zhu and Soricut, 2021]	78.69	Table 2 in [Smith *et al.*, 2022] / Table 1 in [Hasani *et al.*, 2022]
CCNN [Romero *et al.*, 2022]	84.08	Table 2 in [Smith *et al.*, 2022]
S4-v2 (re-impl) [Gu *et al.*, 2022b]	86.53	Table 2 in [Ma *et al.*, 2023]
CDIL [Cheng *et al.*, 2023]	86.78	Table 1 in [Hasani *et al.*, 2022]
S4-v2 / S4 (updated) [Gu *et al.*, 2022b]	**86.82**	Table 2 in [Ma *et al.*, 2023] / Table 10 in [Gu *et al.*, 2022b]
S4-LegS [Gu *et al.*, 2022a]	86.82	Table 2 in [Smith *et al.*, 2022], Table 1 in [Hasani *et al.*, 2022]
S4D-Inv [Gu *et al.*, 2022a]	87.34	Table 1 in [Hasani *et al.*, 2022]
Liquid-S4 / Liquid-S4-PB	89.02	Table 2 in [Smith *et al.*, 2022] / Table 1 in [Hasani *et al.*, 2022]
S5 [Smith *et al.*, 2022]	89.31	Table 2 in [Smith *et al.*, 2022], Table 1 in [Hasani *et al.*, 2022]
MEGA [Ma *et al.*, 2023]	**90.43**	Table 2 in [Ma *et al.*, 2023], Table 2 in [Smith *et al.*, 2022]

LRA – Retrieval

BigBird [Zaheer *et al.*, 2020]	59.29	Table 1 in [Ma *et al.*, 2021], Table 2 in [Ma *et al.*, 2023], Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
Sparse Trans. / Sparse Trans. / Sparse Transformer [Child *et al.*, 2019]	59.59	Table 1 in [Ma *et al.*, 2021] / Table 10 in [Gu *et al.*, 2022b] / Table 1 in [Hasani *et al.*, 2022]
FNet [Lee-Thorp *et al.*, 2021]	59.61	Table 10 in [Gu *et al.*, 2022b], Table 1 in [Hasani *et al.*, 2022]
H-Trans.-1D / H-Transformer-1D [Zhu and Soricut, 2021]	63.99	Table 2 in [Smith *et al.*, 2022] / Table 1 in [Hasani *et al.*, 2022]
Performer in [Choromanski *et al.*, 2020]	78.62	Table 3 in [Xiong *et al.*, 2021]
Reformer [Kitaev *et al.*, 2020]	78.64	Table 1 in [Zhu *et al.*, 2021], Table 3 in [Xiong *et al.*, 2021]
Transformer (re-impl) / XFM (re-impl) [Vaswani *et al.*, 2017]	79.14	Table 1 in [Ma *et al.*, 2021] / Table 2 in [Ma *et al.*, 2023]
Luna-256 [Ma *et al.*, 2021]	79.29	Table 10 in [Gu *et al.*, 2022b], Table 2 in [Smith *et al.*, 2022], Table 1 in [Hasani *et al.*, 2022]
Standard [Vaswani *et al.*, 2017]	79.35	Table 3 in [Xiong *et al.*, 2021]
Linformer [Wang *et al.*, 2020a]	79.37	Table 3 in [Xiong *et al.*, 2021], Table 1 in [Zhu *et al.*, 2021]
Luna-256 [Ma *et al.*, 2021]	**79.56**	Table 1 in [Ma *et al.*, 2021], Table 2 in [Ma *et al.*, 2023]
Nystromformer [Xiong *et al.*, 2021]	79.56	Table 3 in [Xiong *et al.*, 2021], Table 1 in [Hasani *et al.*, 2022], Table 10 in [Gu *et al.*, 2022b]
Nystromformer [Xiong *et al.*, 2021]	81.29	Table 1 in [Zhu *et al.*, 2021]
Performer in [Choromanski *et al.*, 2020]	81.7	Table 1 in [Zhu *et al.*, 2021]
Transformer-LS [Zhu *et al.*, 2021]	81.95	Table 1 in [Zhu *et al.*, 2021]
Full Attention [Vaswani *et al.*, 2017]	82.3	Table 1 in [Zhu *et al.*, 2021]
CDIL [Cheng *et al.*, 2023]	85.36	Table 1 in [Hasani *et al.*, 2022]
DSS [Gupta *et al.*, 2022]	87.6	Table 1 in [Hasani *et al.*, 2022]
S4-v2 / S4 (updated) [Gu *et al.*, 2022b]	**90.9**	Table 2 in [Ma *et al.*, 2023] / Table 10 in [Gu *et al.*, 2022b]
S4-LegS [Gu *et al.*, 2022a]	90.9	Table 2 in [Smith *et al.*, 2022], Table 1 in [Hasani *et al.*, 2022]
S4-v2 (re-impl) [Gu *et al.*, 2022b]	90.94	Table 2 in [Ma *et al.*, 2023]
S4D-Inv [Gu *et al.*, 2022a]	91.09	Table 1 in [Hasani *et al.*, 2022]
Liquid-S4 / Liquid-S4-PB	91.2	Table 2 in [Smith *et al.*, 2022] / Table 1 in [Hasani *et al.*, 2022]
MEGA [Ma *et al.*, 2023]	**91.25**	Table 2 in [Ma *et al.*, 2023], Table 2 in [Smith *et al.*, 2022]
S5 [Smith *et al.*, 2022]	**91.4**	Table 2 in [Smith *et al.*, 2022], Table 1 in [Hasani *et al.*, 2022]


Let’s summarize our observations.

Key Observations

from this Quantitative Review

1. Reducing time and/or memory complexity would inevitably sacrifice a bit non-efficiency performance like accuracy. When the complexities of a set of models were reduced to a certain level, one could claim that the model that achieved the highest accuracy would be the most efficient solution. Past empirical studies unanimously performed efficiency evaluation based on this hypothesis.

Comment: Indirect evaluation and comparison might not be what we need. The real objective should be time and space complexity when the focus is model efficiency. Should we compare their marginal differences on accuracy-based performance when the complexity of the models was not reported or analyzed?

Key Observations (cont.)

from this Quantitative Review

2. Most empirical studies compared their proposed model against others on multiple tasks, and usually claimed theirs is the best one on all of them. However, our review identifies different winning approaches for different tasks and even different datasets. It fixes the one-sided understanding that researchers would have from learning only one or a few empirical studies.

3. More than half of the results were reported by at least two sources. However, it is impossible to tell from the papers whether the numbers were reproduced/confirmed or just re-used from previous work. Inconsistent results were found on almost every task, caused by various settings of hyperparameters (e.g., model sizes, configurations) and reproductions.

Conclusions

from this Quantitative Review

1. This literature review on **LM efficiency** offered a set of integrated comparative results that would not be observed without such an effort. It covered both Transformer models and the emergent state space models.
2. Researchers are suggested to report the performance of baselines properly, write clearly if the numbers in experimental results were re-used from prior work or reproduced, and report and analyze any inconsistent results that are identified compared with related studies.
3. Researchers in this field need public leaderboards. We need to think about evaluation metrics for the leaderboards.

Faculty

Meng Jiang
Associate Professor

PhD Cand.

Wenhao Yu
Bloomberg Fellowship
ICLR'23, EMNLP'22 * 2, ACL'22 * 2,
EMNLP'21 * 2, NAACL'21, WWW'20 * 2
LLM and open-domain QA

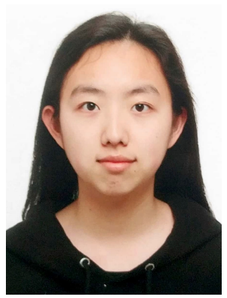
Qingkai Zeng
KDD'21, EMNLP'20
Information Extraction
Taxonomy Construction

PhD Student

Zhihan Zhang
TACL'23, EMNLP'22
LLM and open-domain QA
Instruction Tuning

Noah Ziems
ACL'23
LLM and open-domain QA
Controlled Text Generation

Mengxia Yu
CSE Select Fellowship
LLM for Education
Comparative Reasoning



Lingbo Tong
Psychology-CSE Joint PhD
LLM for Mental Health
Knowledge Graph

Hy Dang
LLM for Mental Health
Query Expansion

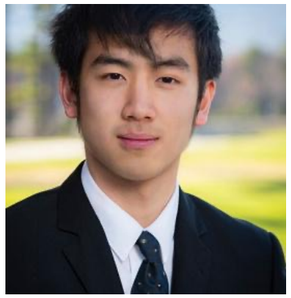
Bang Nguyen
LLM for Education
Query/Question Generation

Zhaoxuan Tan
LLM for User Profiling

Mengzhao Jia
LLM for User Profiling

Gang Liu
KDD'22
KDD'23

Eric Inae
Dean's Fellowship
GNN for Material Science



Zheyuan Liu

Welcome to visit our lab!

http://www.meng-jiang.com/lab.html

https://github.com/DM2-ND

mjiang2@nd.edu

UNIVERSITY OF NOTRE DAME

College of Engineering