SpaceByte: Towards Deleting Tokenization from Large Language Modeling

Kevin Slagle

Rice University
(now at Magic)

Dec 11, 2024

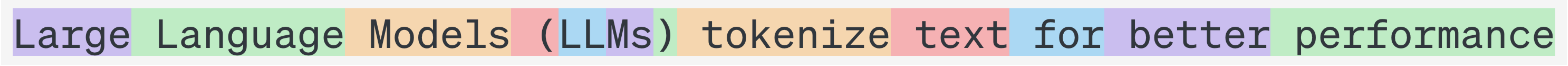
Motivation

Tokens

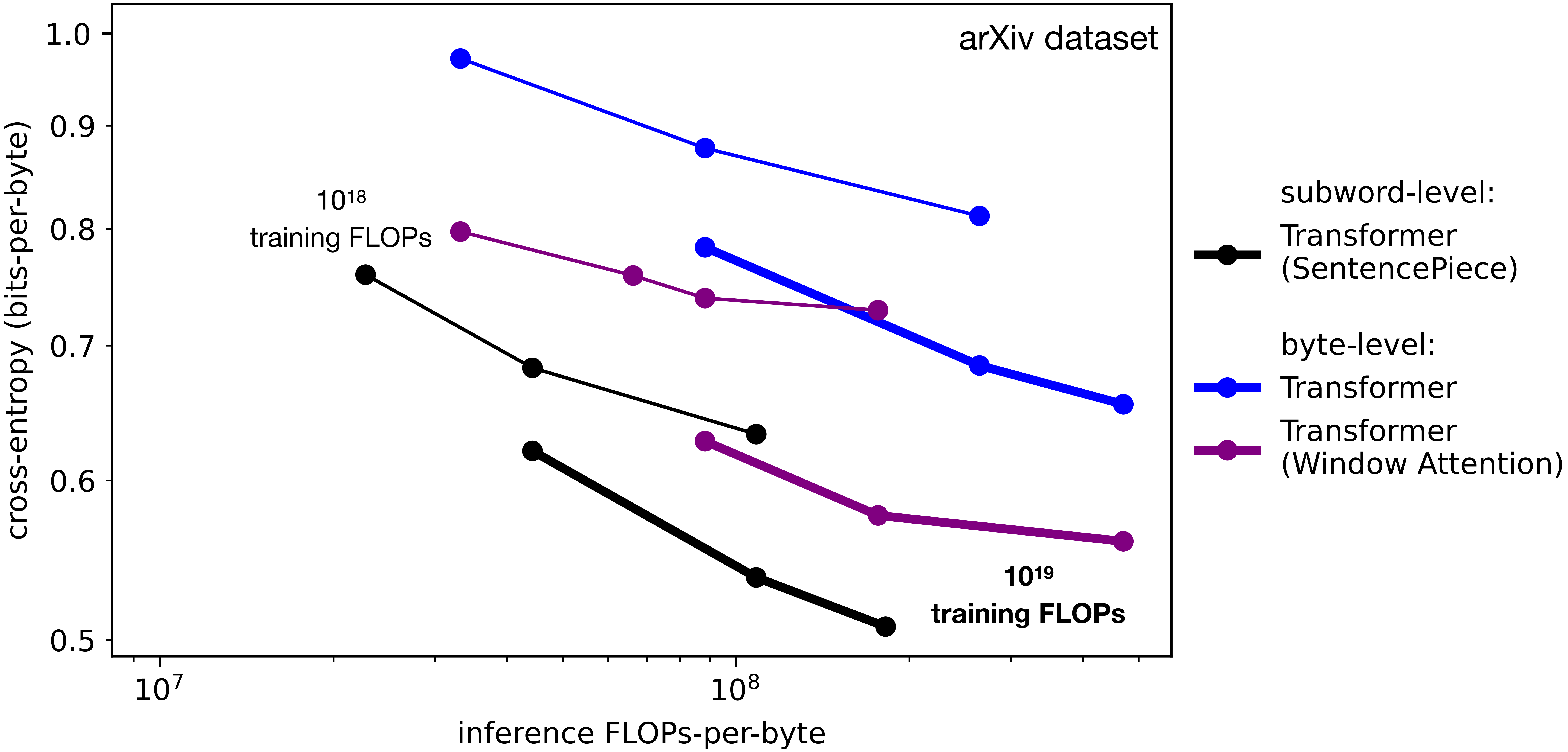
12

Characters

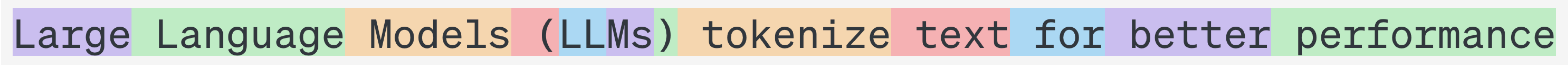
65

- 

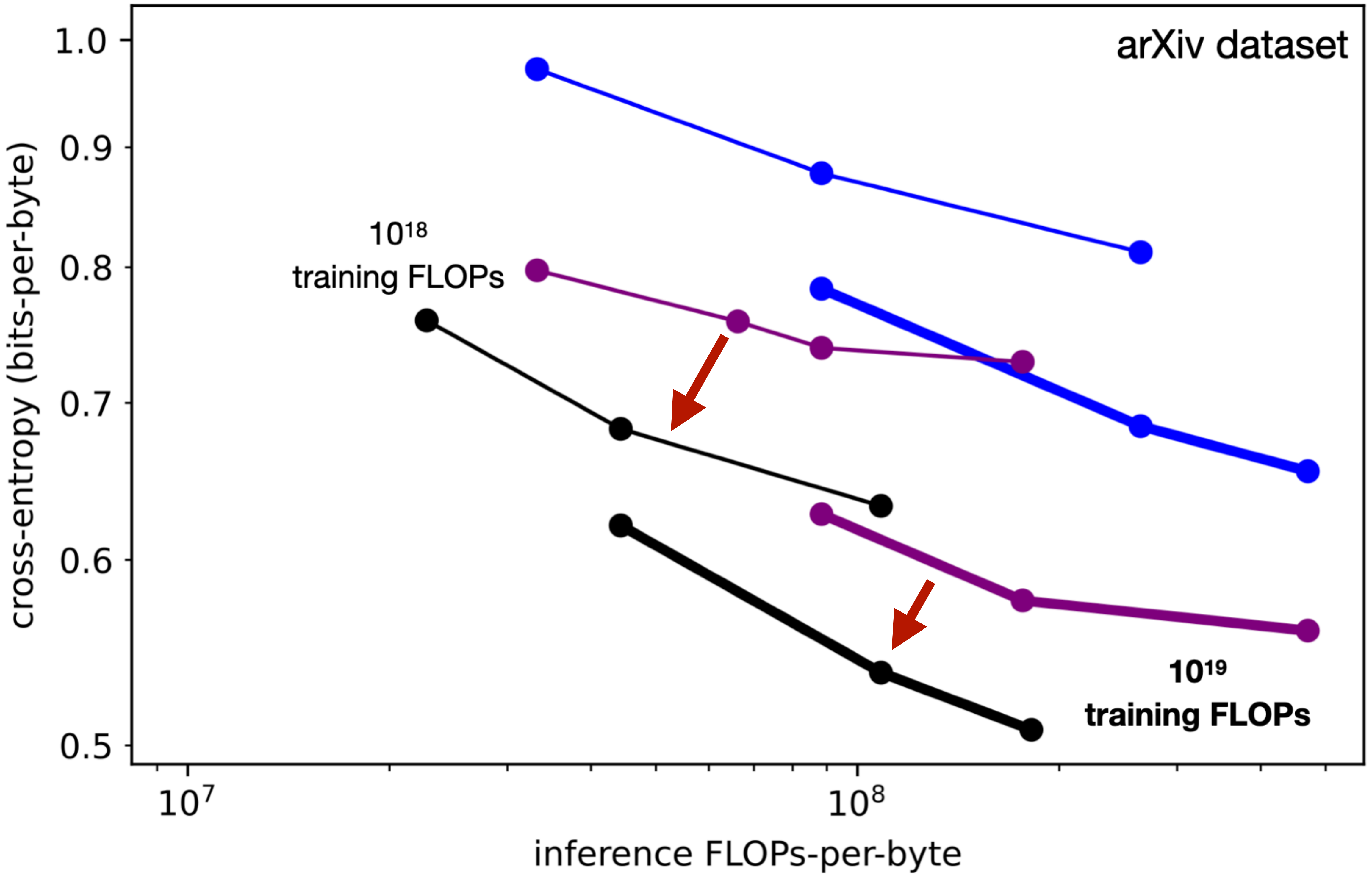
  Large Language Models (LLMs) tokenize text for better performance



Motivation

- 
  Large Language Models (LLMs) tokenize text for better performance

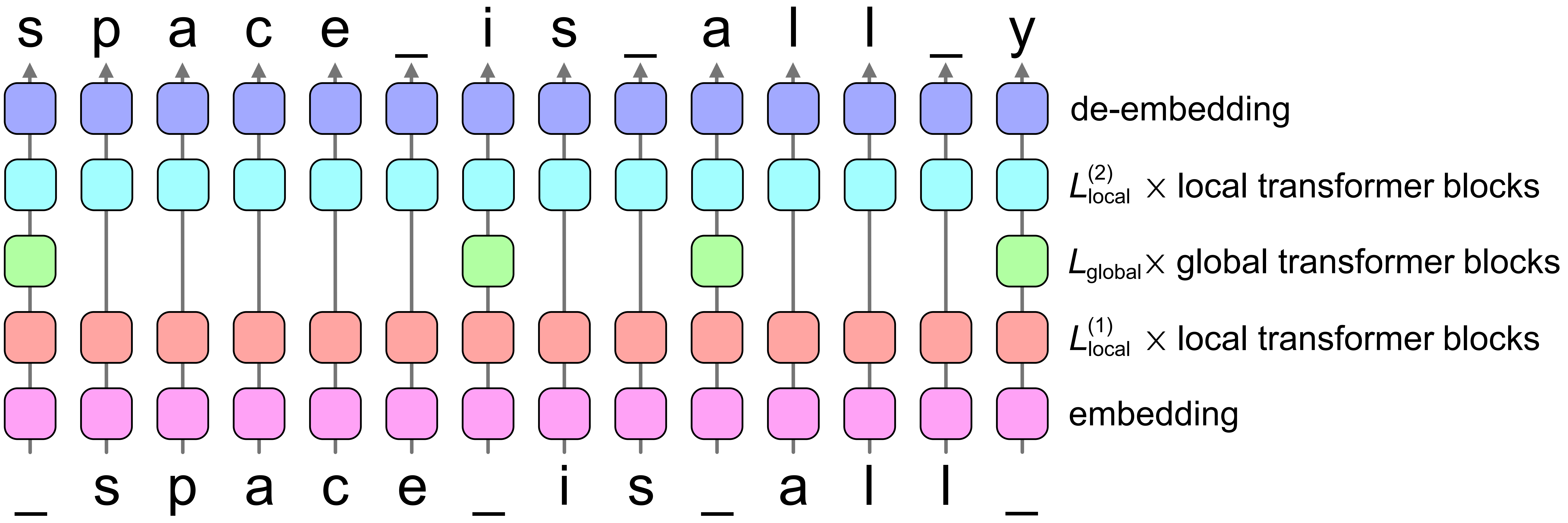
- We would like to avoid tokenization
  - Less modeling complexity
  - Less adversarial vulnerability
  - Better character-level performance



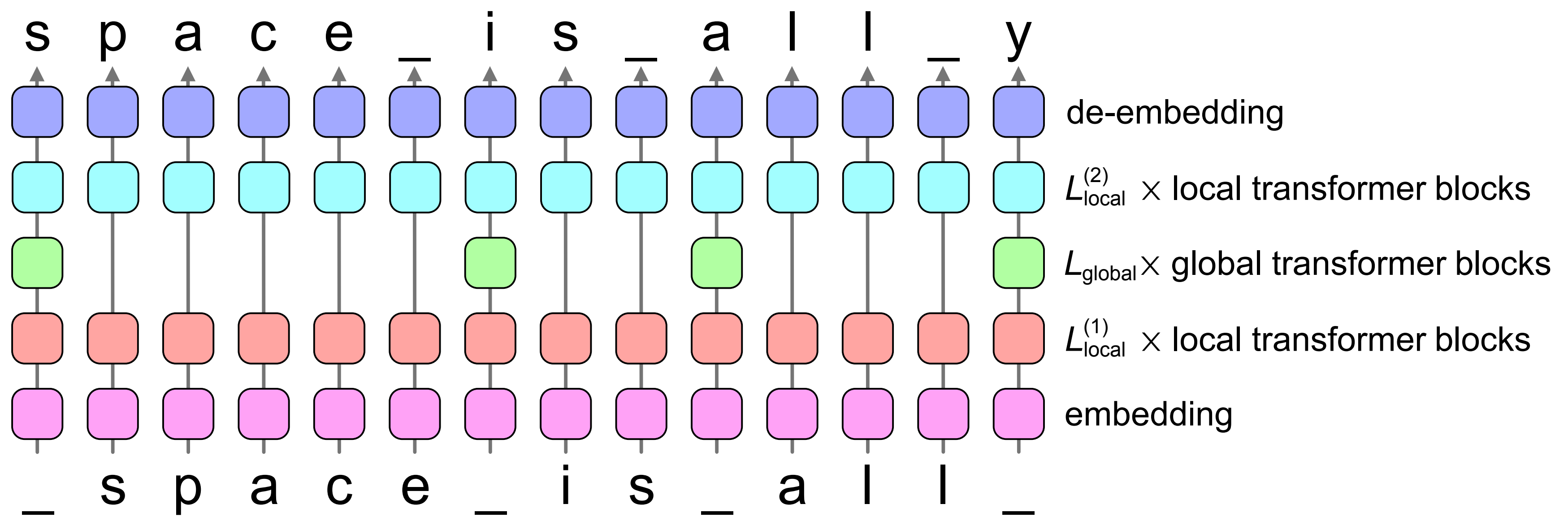
- Requires closing the performance gap

SpaceByte Model

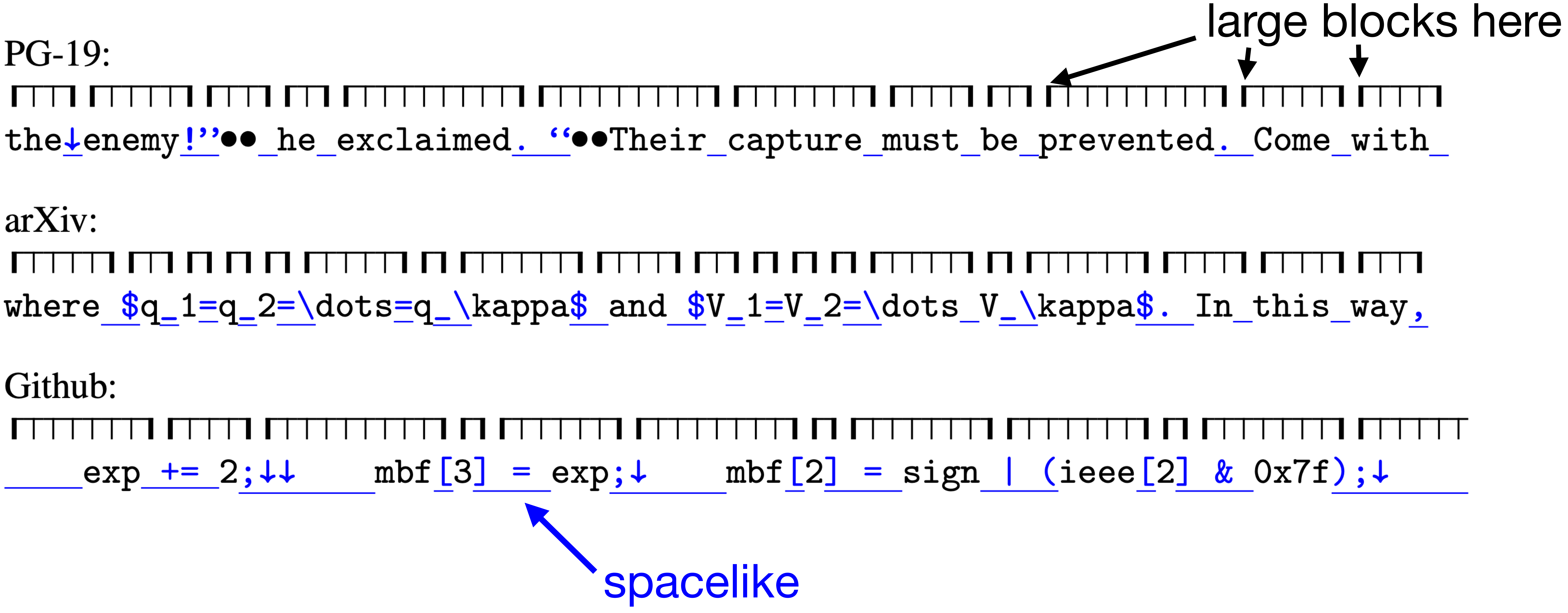
- Insert large transformer blocks after “spacelike” characters
  - spacelike ~ not a letter or number


s p a c e _ i s _ a l l _ y
de-embedding
$L_{\text{local}}^{(2)}$ × local transformer blocks
L_{global} × global transformer blocks
$L_{\text{local}}^{(1)}$ × local transformer blocks
embedding
_ s p a c e _ i s _ a l l _

SpaceByte Model


s p a c e _ i s _ a l l _ y
de-embedding
$L_{\text{local}}^{(2)}$ × local transformer blocks
L_{global} × global transformer blocks
$L_{\text{local}}^{(1)}$ × local transformer blocks
embedding
_ s p a c e _ i s _ a l l _

- Insert large transformer blocks after “spacelike” characters



PG-19:

```
the↓enemy!’’●●_he_exclaimed._‘‘●●Their_capture_must_be_prevented._Come_with_
```

arXiv:

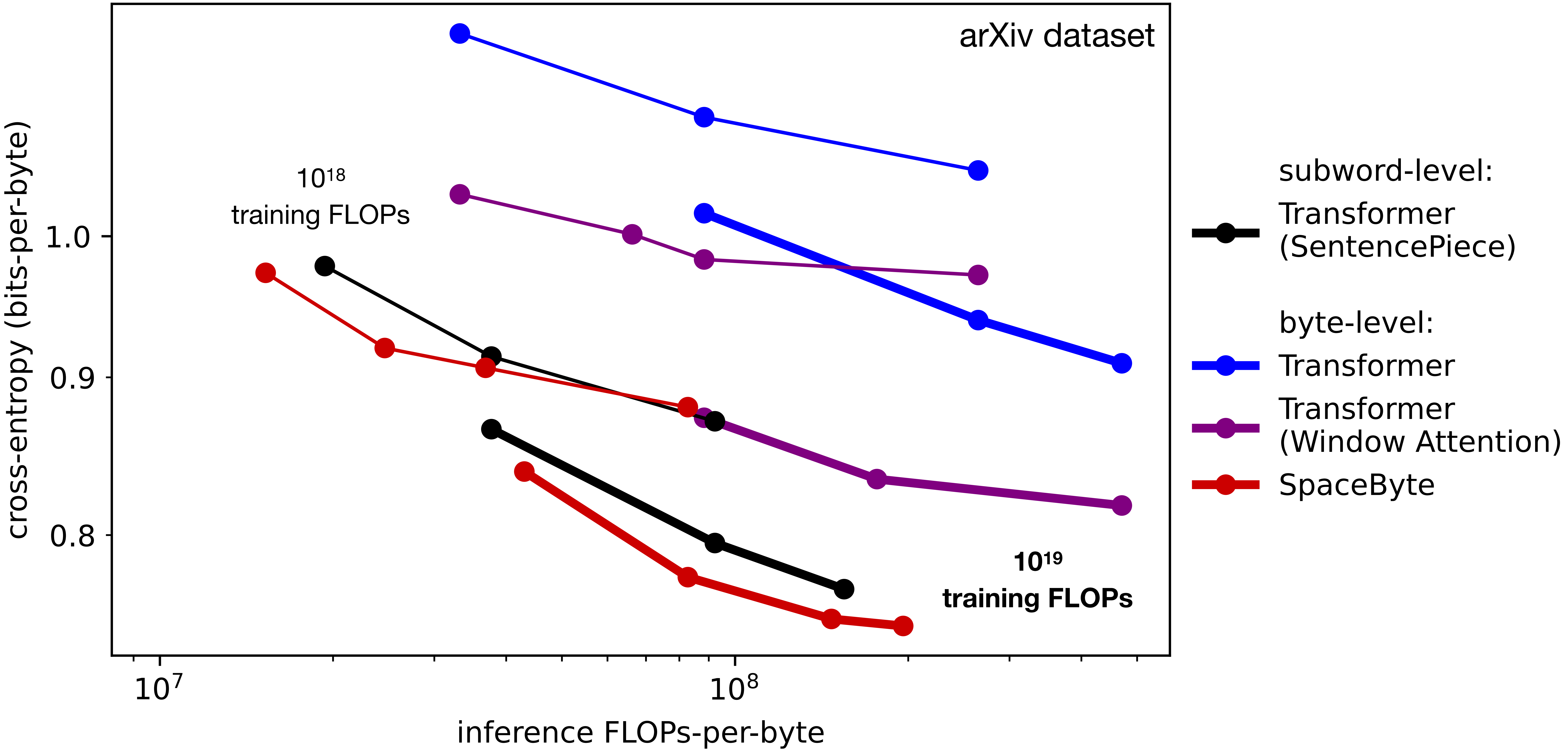
```
where_$q_1=q_2=\dots=q_\kappa$_and_$V_1=V_2=\dots_V_\kappa$._In_this_way,
```

Github:

```
____exp_+=_2;↓↓_______mbf[3]_=_exp;↓_______mbf[2]_=_sign_|_(ieee[2]_&_0x7f);↓____
```

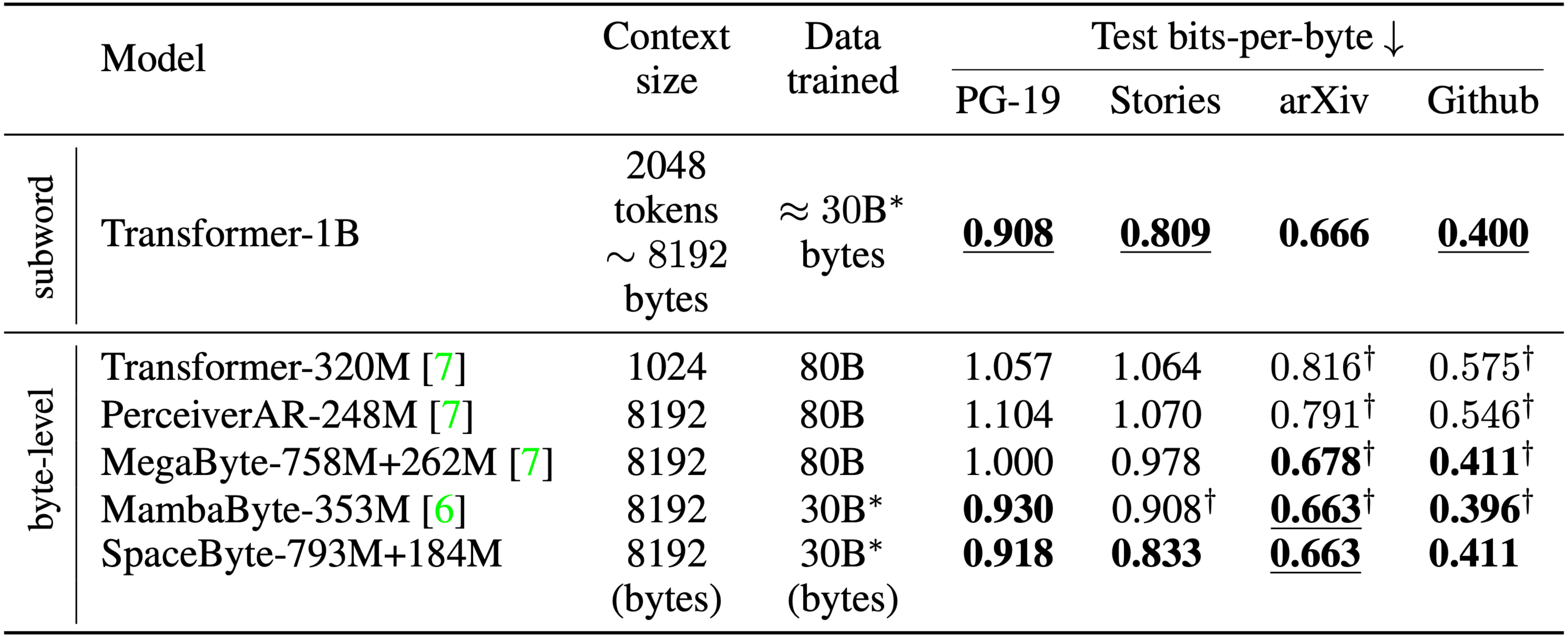
Scaling Analysis

- SpaceByte performs slightly better than subword models!



Comparison with Other Works

- SpaceByte is competitive with subword Transformer and MambaByte

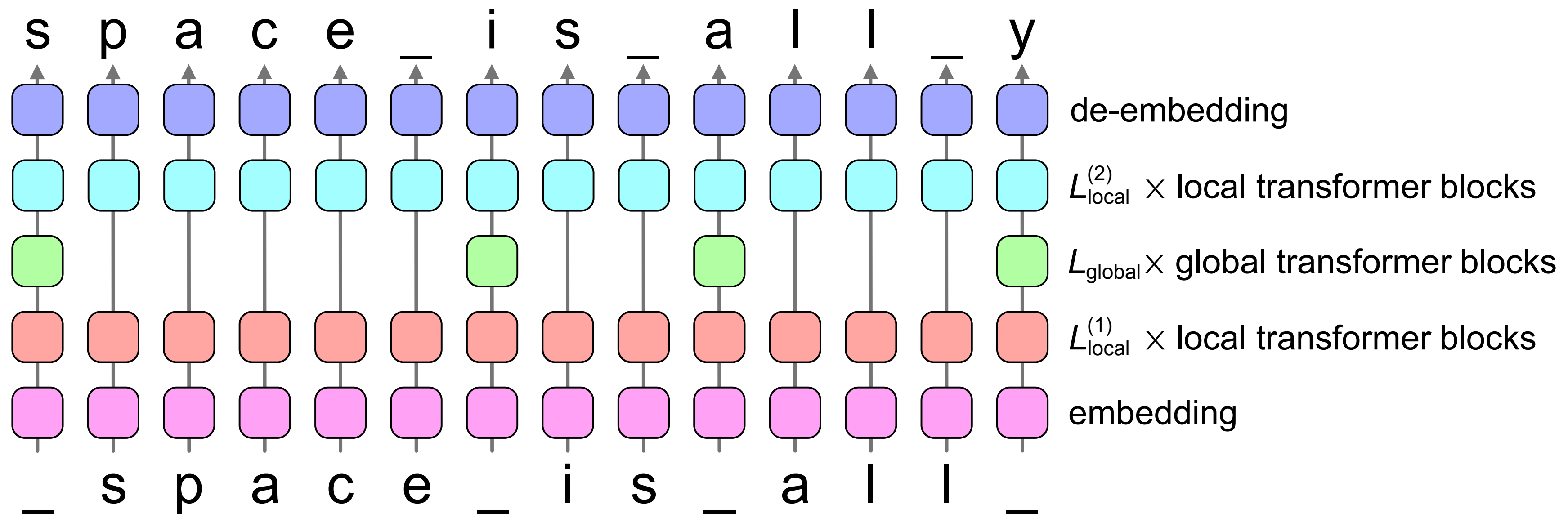


	Model	Context size	Data trained	Test bits-per-byte ↓			
				PG-19	Stories	arXiv	Github
subword	Transformer-1B	2048 tokens $\sim$ 8192 bytes	≈ 30B* bytes	**0.908**	**0.809**	**0.666**	**0.400**
byte-level	Transformer-320M [7]	1024	80B	1.057	1.064	0.816†	0.575†
	PerceiverAR-248M [7]	8192	80B	1.104	1.070	0.791†	0.546†
	MegaByte-758M+262M [7]	8192	80B	1.000	0.978	**0.678**†	**0.411**†
	MambaByte-353M [6]	8192	30B*	**0.930**	0.908†	**0.663**†	**0.396**†
	SpaceByte-793M+184M	8192	30B*	**0.918**	**0.833**	**0.663**	**0.411**
		(bytes)	(bytes)				

6.5x10^{19} training FLOPs

Conclusion

- We introduce SpaceByte:
  - A multi-scale transformer architecture
  - Models byte-level language (rather than tokens) w/o performance penalty


s p a c e _ i s _ a l l _ y
de-embedding
$L_{\text{local}}^{(2)} \times$ local transformer blocks
$L_{\text{global}} \times$ global transformer blocks
$L_{\text{local}}^{(1)} \times$ local transformer blocks
embedding
_ s p a c e _ i s _ a l l _

- Limitations and future work:
  - Languages that don’t use space characters (e.g. Chinese)?
  - Batched inference is more complicated
  - Multiscale modeling at larger scales?
    - E.g. sentence-level rather than world-level