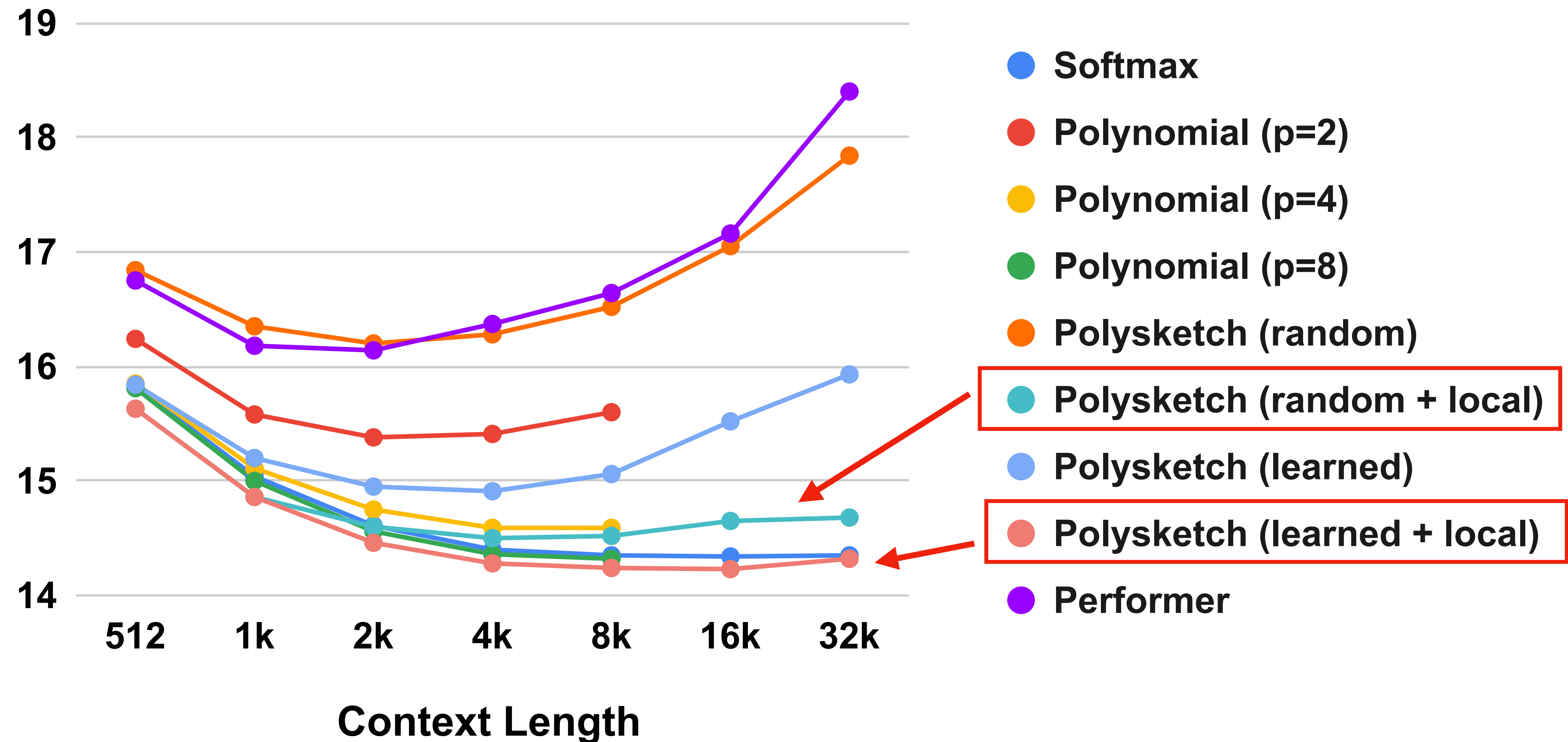


Model Perplexities

Perplexities on Wiki-40B



Down Stream Tasks

	C4	HellaSwag		PIQA		Physics	
	Perplexity ↓	0-shot ↑	5-shot ↑	0-shot ↑	5-shot ↑	0-shot ↑	5-shot ↑
GPT-2 Small style, 100M-scale, 12 layers default, Context Length 8192, 125k training steps							
Softmax	17.81	30.2	27.8	64.6	63.2	27.5	27.5
Polynomial (degree 4)	18.18	28.6	<u>28.4</u>	64.2	65.0	<u>27.5</u>	<u>31.0</u>
Polynomial (degree 8)	<u>17.77</u>	29.8	<u>29.8</u>	62.2	<u>64.0</u>	23.1	26.2
Polysketch (learned, r = 32)	19.09	28.0	28.4	60.6	62.0	28.3	27.5
Polysketch (learned, 13 layers, r = 32)	19.50	28.4	29.0	61.6	<u>64.6</u>	27.9	<u>33.1</u>
Polysketch (learned + local, r = 32)	18.04	29.0	29.2	63.4	62.8	26.6	35.8
Polysketch (learned + local, 13 layers, r = 32)	<u>17.72</u>	31.2	<u>30.4</u>	64.8	<u>64.6</u>	<u>27.9</u>	<u>31.8</u>
GPT-2 Large style, 700M-scale, 36 layers default, Context Length 2048, 125k training steps							
Softmax	12.71	40.2	40.2	68.8	71.4	34.4	24.4
Polynomial (degree 4)	12.82	40.0	40.6	67.8	66.6	31.8	<u>31.4</u>
Polynomial (degree 8)	12.85	40.0	39.8	66.8	70.4	<u>34.4</u>	<u>29.6</u>
Polysketch (learned, 39 layers, r = 32)	12.98	39.4	<u>40.4</u>	68.6	67.6	33.6	27.0
Polysketch (learned + local, 39 layers, r = 32)	12.74	39.6	40.6	66.8	69.4	<u>35.3</u>	<u>31.8</u>