

Comparisons

	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
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	<u>40.6</u>	67.8	66.6	31.8	<u>31.4</u>
Polynomial (degree 8)	12.85	40.0	<u>39.8</u>	66.8	70.4	<u>34.4</u>	<u>29.6</u>

Feature map for Polynomials