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	28.4	64.2	**65.0**	27.5	31.0
Polynomial (degree 8)	17.77	29.8	29.8	62.2	64.0	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	**40.6**	67.8	66.6	31.8	31.4
Polynomial (degree 8)	12.85	40.0	39.8	66.8	70.4	34.4	29.6

Feature map for Polynomials

- A finite dimensional φ such that $\langle \varphi(q), \varphi(k) \rangle = \langle q, k \rangle^p$?
  - $\varphi : x \mapsto x^{\otimes p}$
  - If $x \in \mathbb{R}^h$, then $x^{\otimes p} \in \mathbb{R}^{h^p}$
  - $(x^{\otimes p})_{(i_1, i_2, \ldots, i_p)} = x_{i_1} \cdot x_{i_2} \cdot \cdots \cdot x_{i_p}$