Table R1: Relative error and perplexity on Wikitext2 after quantizing the weights of DeepSeek-MoE-16B using the RTN quantization algorithm.

Bitwidth	**2**	**3**	**4**	**5**	**6**	**8**	**Fp16**
Mean Relative Error	92.67%	38.01%	14.26%	7.13%	3.56%	0.90%	0.00%
Perplexity ↓	5e8	10.65	7.151	6.629	6.539	6.51	6.50

Table R2:

τ	**1.0**	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**
DeepSeek-MoE-16B	39.82	39.89	40.01	39.89	39.69	39.71
QwenMoE-14B						
Mixtral-8x7B						