Table 1: Comprehensive evaluation of 2-bit weight quantization (W2A16) on LLaMA-2 models.

Method	WinoG	PIQA	HellaS	ARC-e	ARC-c	MM
LLaMA2-7B (FP16)	69.06	78.07	57.14	76.30	43.34	41.8
GPTQ	48.93	57.13	28.15	32.11	20.22	22.9
AWQ	49.57	52.39	0.11	38.89	20.73	22.9
OmniQuant	51.54	57.40	30.11	38.89	20.73	22.9
QuIP	51.07	59.25	30.11	38.89	20.73	22.9
ButterflyQuant	62.27	68.97	48.43	62.58	29.86	26.6
LLaMA2-13B (FP16)	72.14	79.11	60.04	79.46	48.46	52.1
GPTQ	52.09	62.24	34.80	42.59	21.25	23.0
AWQ	49.57	53.26	25.81	23.04	23.04	26.8
OmniQuant	52.17	62.89	40.16	48.23	24.66	22.9
QuIP	55.72	65.45	39.65	51.56	25.85	23.7