Table 1: Perplexity comparisons of the OPT models and LLaMA models with 4-bit quantization on WikiText-2. We set the group size as the length of rows for OPT models and 128 for LLaMA models following baselines for fair comparisons.

Method	Bits	OPT							LLaMA			
		125M	350M	1.3B	2.7B	6.7B	13B	30B	1-7B	2-7B	2-13B	3-8B
OPTQ	4	31.12	24.24	15.47	12.87	11.39	10.31	9.63	6.22	5.69	4.98	7.63
LUT-GEMM	4	31.93	24.09	16.15	13.34	12.09	10.40	9.99	5.94	5.78	5.06	6.85
AWQ	4	31.66	7.4e3 (outlier)	15.22	13.19	11.23	-	-	5.78	5.60	4.97	-
Ours (Acc.)	4	**28.72**	**21.59**	**14.98**	**12.65**	**10.95**	**10.20**	**9.63**	**5.76**	**5.58**	**4.96**	**6.46**

Table 2: Perplexity comparisons between ShiftAddLLM and OmniQuant using OPT models and LLaMA models on WikiText-2. The group size is set as the length of rows for OPT models and 128 for LLaMA models following baselines.

Method	Bits	OPT								LLaMA-2		
		125M	350M	1.3B	2.7B	6.7B	13B	30B	66B	7B	13B	70B
OmniQuant	4	29.45	23.19	15.04	12.76	11.03	10.30	9.65	-	5.58	4.95	-
Ours (Acc.)	4	**28.72**	**21.59**	**14.98**	**12.65**	**10.95**	**10.20**	**9.63**	-	5.58	4.96	-
OmniQuant	3	35.66	28.2	16.68	13.8	11.65	10.87	10.00	9.83	6.03	5.28	3.78
Ours (Acc.)	3	**31.29**	**24.24**	**21.53**	**13.68**	**11.18**	**10.39**	**9.63**	**9.43**	**5.89**	**5.16**	**3.64**
OmniQuant	2	311.39	186.9	484.51	1.1e6	9.6e5	3.6e4	9.3e3	5.2e3	11.06	8.26	6.55
Ours (Acc.)	2	**51.15**	**40.24**	**29.03**	**20.78**	**13.78**	**12.17**	**10.67**	**10.33**	**8.51**	**6.77**	**4.72**

Table 3: Perplexity comparisons between ShiftAddLLM and FlexRound. The group size of FlexRound is set as the length of rows following the paper.

Method	Bits	LLaMA-2		
		7B	13B	70B
FlexRound	4	5.83	5.01	-
Ours (Acc.)	4	5.58	4.96	-
FlexRound	3	6.34	5.59	3.92
Ours (Acc.)	3	**5.89**	**5.16**	**3.64**

Table 4: Accuracy comparisons on eight downstream tasks for OPT-66B and LLaMA-2-70B with additional MMLU.

Models	Methods	Bits	ARC_C	ARC_E	Copa	BoolQ	PIQA	Storycloze	RTE	MMLU	Mean
OPT-66B	Float-Point	16	37.20	71.25	86	69.82	78.67	77.47	60.65	25.89±0.37	63.37
	OPTQ	3	24.66	48.86	70	52.05	64.47	67.09	53.07	23.98±0.36	50.52
	LUT-GEMM	3	24.15	51.85	81	53.52	61.97	60.60	48.74	23.73±0.36	50.70
	Ours (Acc.)	3	**35.24**	**70.88**	**87**	**72.45**	**77.64**	**77.15**	**63.18**	**27.56±0.38**	**63.89**
LLaMA-2-70B	Float-Point	16	49.57	76.14	90	82.57	80.79	78.61	68.23	65.24±0.37	72.33
	OPTQ	3	45.82	76.34	90	81.74	79.71	77.34	67.51	60.14±0.36	68.31
	LUT-GEMM	3	47.70	76.42	89	80.31	80.20	77.78	68.59	-	-
	Ours (Acc.)	3	**48.38**	**77.06**	**93**	**84.25**	**80.47**	**78.49**	**75.09**	**62.33±0.38**	**74.88**

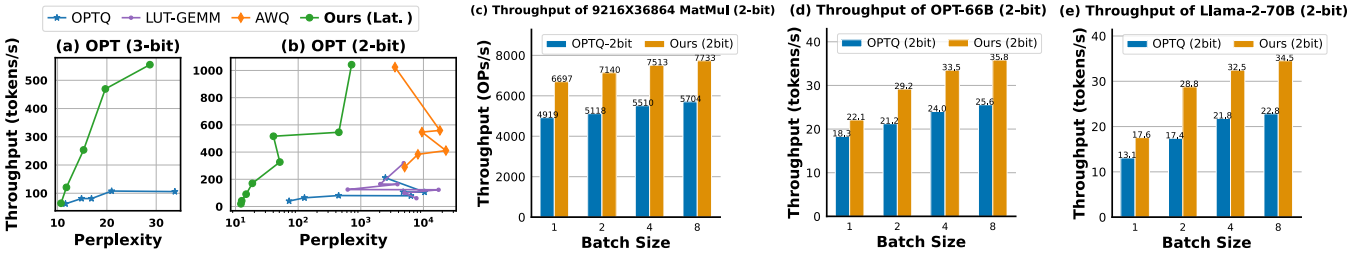

Figure 1: (a-b): Accuracy-throughput tradeoff comparisons among ShiftAddLLM, OPTQ, LUT-GEMM, and AWQ at a batch size of 8. (c) Kernel throughput evaluation under batch sizes of 1, 2, 4, and 8. (d) LLaMA-2-70B end-to-end model throughput evaluation under batch sizes of 1, 2, 4, and 8. (e) OPT-66B end-to-end model throughput evaluation under batch sizes of 1, 2, 4, and 8.

Table 5: More tests on various optimization objectives.

PPL ↓	OPT (3-bit)			LLaMA (3-bit)		
	125m	1.3b	6.7b	1-7B	2-7B	2-13B
Wei. Obj.	60.00	49.10	17.44	7.18	7.02	5.89
Act. Obj.	53.85	20.97	14.86	8.81	6.43	5.48
Wei. + Act.	205.45	142.68	52.66	10.15	6.89	6.79
Ours (Multi-Obj.)	**31.29**	**21.53**	**11.18**	**6.04**	**5.89**	**5.16**

Table 6: Perplexity and correlation results of our mixed bit allocation.

Methods	Bits	OPT			LLaMA			Gemma
		125M	1.3B	13B	2-7B	2-13B	3-8B	2B
Correlation(τ)		0.910	0.905	0.915	0.931	0.929	0.897	-
Ours (Lat.)	2	712.55	40.28	14.76	9.58	12.57	34.40	16.52 (3 bits)
Ours (Mixed)	2.2	435.84	27.37	13.79	8.97	8.16	29.72	13.89

Input: Can you cook an egg only using durect sunlight in any place of our solar system ?

FP model: Can you make a solar oven that can cook a whole egg? I'm not sure if you can make an omelet in a regular solar cooker, but you could make scrambled eggs. You can't cook eggs in the sun. You can only cook them in an environment that is hotter than the boiling point of water. The sun is not hot enough to cook anything.	**OPTQ :** Can you use the sun to cook eggs? I'm not sure if you can cook the egg in the Sun, but you could use a solar oven to boil water. You can use solar energy to make a hot water bottle. You can also use it to heat your home.	**LUT-GEMM:** in the sun ? in the moon ? on the planets ? or in a star ? and what about the other planes of the universe ? can we cook in them too ?	**Ours:** I'm not sure what you mean by "direct sun light". If you are asking if you can cook eggs in the sun, then the answer is yes. The sun is a source of heat and light. It is not a cooking device. **You can use the heat from the Sun to cook food, but you will need to use a device to concentrate the light and heat.** The Sun is too far away to be used as a direct cooker.

Figure 2: Responses of different quantization methods to the same input using LLaMA-2-7B with 3-bit quantization. We color the text to show the correct or wrong responses.