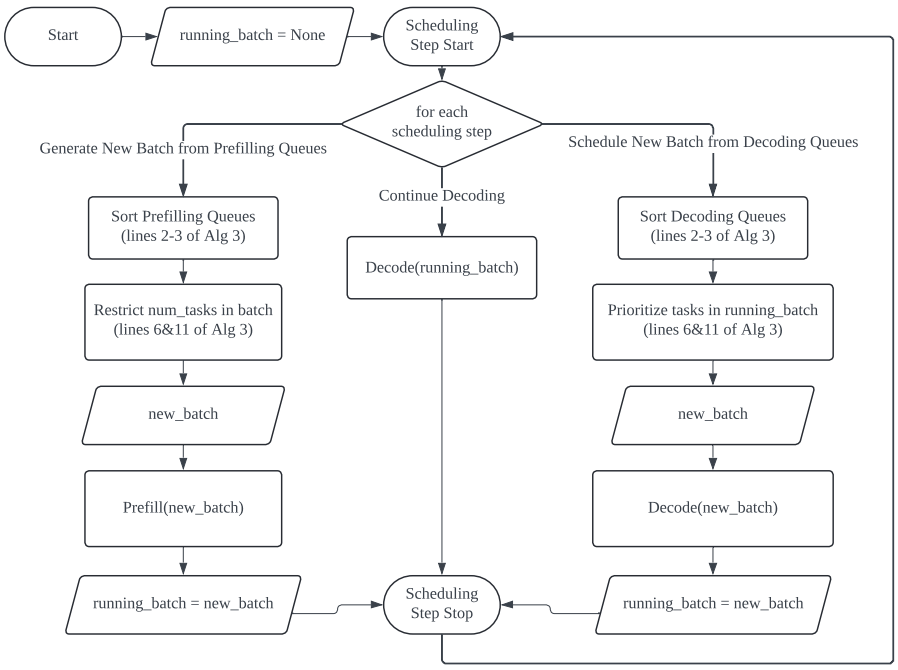

Figure A: Flow chart of our multi-task scheduling strategy.

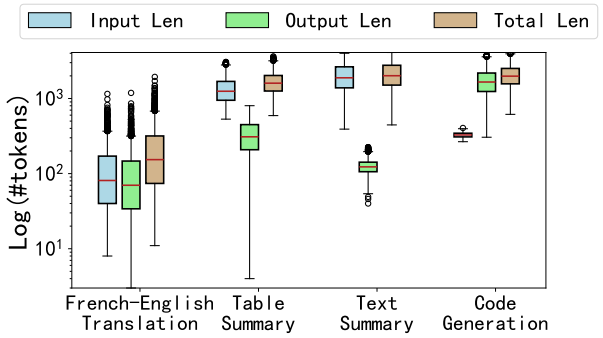

Figure B: Length distributions of different tasks in Log scale.

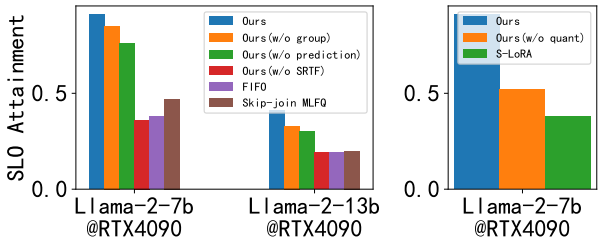

Figure C: Left: Ablation studies of scheduling strategies on *LoRA-Inlaid* (100 tasks). Right: System comparison (Llama-2-13B is not shown due to OOM of the other two methods).

Algorithm A Routines of MLAWQ and AWQ$_{tweaked}$ to quantize on layer in multi-task scenarios.

Input: $\{\mathbf{X}^t\}_{t=1}^T$, ratio_search_space ▷ The inputs of different tasks of this layer

$\mathbf{M}^t \leftarrow (\mathbf{W}+\mathbf{B}^t\mathbf{A}^t)\mathbf{X}^t$ $\mathbf{M}_{\text{mixed}} \leftarrow \sum_1^T \mathbf{W}\mathbf{X}^t$ ▷ Forward to get monitoring matrix

$\mathbf{W}_{\text{mean}} \leftarrow \mathbf{W}.\text{mean}(0)$

$\mathbf{X}^t_{\text{mean}} \leftarrow \mathbf{X}^t.\text{mean}(0)$ for $\forall t \in [1,T]$ $\mathbf{X}_{\text{mean}} \leftarrow \{\mathbf{X}^t\}_{t=1}^T.\text{mean}(0)$ ▷ Aggregate Info: Per-channel mean of $\mathbf{X}$

best_s $\leftarrow$ None

min_errors $\leftarrow \infty \cdot \mathbf{1}_{\dim(X^t.\text{shape}[-1])}$ min_error $\leftarrow \infty$ ▷ Initialize minimum error(s)

for ratio **in** ratio_search_space **do** ▷ Aggregate Info: Calculate best_s

$s_t \leftarrow \left(\frac{\mathbf{X}^t_{\text{mean}}.\text{pow}(\text{ratio})}{\mathbf{W}_{\text{mean}}.\text{pow}(1-\text{ratio})}\right)$ for $\forall t \in [1,T]$ $s \leftarrow \left(\frac{\mathbf{X}_{\text{mean}}.\text{pow}(\text{ratio})}{\mathbf{W}_{\text{mean}}.\text{pow}(1-\text{ratio})}\right)$ ▷ Calculate s

$\mathbf{W}^t_{\text{scaled}} \leftarrow \mathbf{W}\cdot s_t$ $\mathbf{W}_{\text{scaled}} \leftarrow \mathbf{W}\cdot s$ ▷ Scale $\mathbf{W}$

$\mathbf{X}^t_{\text{scaled}} \leftarrow \mathbf{X}^t/s_t$ $\mathbf{X}_{\text{scaled}} \leftarrow \{\mathbf{X}^t\}_{t=1}^T/s$ ▷ Scale $\mathbf{X}$

$\text{errs}_t \leftarrow \|\mathbf{M}^t - (\alpha\,(\text{round}\,(\text{clamp}\,(\mathbf{W}^t_{\text{scaled}}/\alpha, \text{min\_val}, \text{max\_val})))+\mathbf{B}^t\mathbf{A}^t)\mathbf{X}^t_{\text{scaled}}\|_{2,\text{columns}}$

$\text{err} \leftarrow \|\mathbf{M}_{\text{mixed}} - \alpha\,(\text{round}\,(\text{clamp}\,(\mathbf{W}_{\text{scaled}}/\alpha, \text{min\_val}, \text{max\_val})))\,\mathbf{X}_{\text{scaled}}\|$ ▷ Use pseudo quantized W run forward to cal quant error of this ratio, where α is the scale factor of pseudo quant

$\text{min\_errs}[j], \text{best\_s}[j] \leftarrow \min_t(\text{errs}_t[j]), s_{\arg\min_t(\text{errs}_t[j])}[j] \;\; \forall\, j \in \mathbf{X}^t.\text{shape}[-1], \forall t \in [1,T]$

$\text{min\_err} \leftarrow \min(\text{min\_err}, \text{err}), \quad \text{best\_s} \leftarrow (\text{err} < \text{min\_err})\,?\,s\,:\,\text{best\_s}$

▷ Aggregate the min error to get best_s

$\mathbf{W}_{\text{modified}} \leftarrow \mathbf{W}\cdot \text{best\_s}$ ▷ Modify weight

$\mathbf{W}_{\text{quant}} \leftarrow \text{quantize}(\mathbf{W}_{\text{modified}})$ ▷ Quantize modified weight

Return $\mathbf{W}_{\text{quant}}$

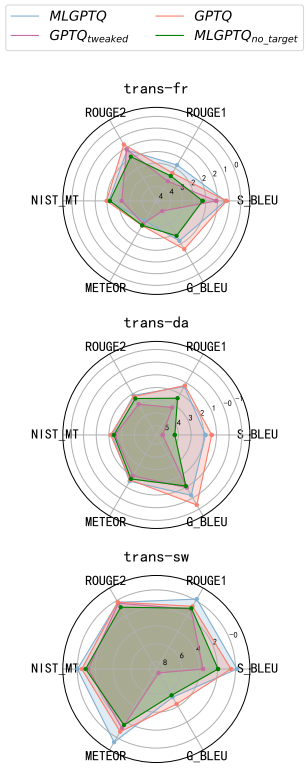

Figure D: More results of effectiveness anatomy.

Table A: Model quality of different approaches under different tasks (std-dev given in parentheses). The rightmost three datasets are added to enhance the diversity of experiments. GPTQ and AWQ are in gray background color since the quantized models produced by them cannot be shared across different tasks. We mark the best multi-task quantization approaches (i.e., the best among MLGPTQ, GPTQ$_{\text{tweaked}}$, AWQ$_{\text{tweaked}}$, and RTN) in bold.

Dataset / Metric	trans-fr SacreBLEU	trans-cs SacreBLEU	trans-id SacreBLEU	trans-nl SacreBLEU	trans-da SacreBLEU	trans-sw SacreBLEU	QTsum ROUGE-1	xlsum ROUGE-1	tiny-codes ROUGE-1	GSM8k Accuracy (%)	Medical_MMLU Accuracy (%)	malicious-600k Accuracy (%)
Unquantized	34.45 (0.01)	31.89 (0.02)	33.94 (0.01)	30.94 (0.01)	35.04 (0)	31.14 (0.01)	49.38 (0)	41.28 (0.01)	31.72 (0.02)	32.14 (0)	80.9 (0)	37.44 (0)
MLGPTQ (4-bit)	**34.05 (0.02)**	**31.16 (0.02)**	**33.63 (0.01)**	**30.73 (0.04)**	**34.39 (0.01)**	**31.20 (0.01)**	**49.02 (0)**	**40.96 (0.02)**	**30.85 (0.05)**	**31.62 (0)**	**76.7 (0.68)**	**36.44 (0.6)**
GPTQ$_{\text{tweaked}}$ (4-bit)	33.91 (0.02)	28.95 (0.21)	32.88 (0.08)	30.48 (0.07)	33.47 (0.04)	28.94 (0.06)	48.23 (0.02)	39.77 (0.06)	29.25 (0.10)	31.51 (0)	74.81 (2.53)	35.05 (1.15)
AWQ$_{\text{tweaked}}$ (4-bit)	33.88 (0.04)	29.45 (0.11)	33.01 (0.06)	29.99 (0.09)	33.34 (0.11)	30.11 (0.07)	47.96 (0.03)	40.12 (0.08)	30.23 (0.07)	30.51 (0.01)	75.42 (1.13)	35.68 (0.33)
RTN (4-bit)	33.79 (0)	29.64 (0.01)	32.96 (0.01)	30.33 (0)	33.96 (0)	30.46 (0.02)	47.54 (0.01)	40.27 (0.02)	30.63 (0.02)	31.01 (0)	76.15 (0)	33.78 (0)
GPTQ (4-bit)	34.07 (0.02)	31.19 (0.03)	33.79 (0.02)	30.86 (0.15)	34.57 (0.02)	31.08 (0.08)	49.26 (0.02)	40.89 (0.06)	30.92 (0.06)	31.35 (0)	76.7 (2.66)	36.25 (0.44)
AWQ (4-bit)	34.17 (0.03)	31.19 (0.05)	33.72 (0.07)	30.69 (0.08)	34.21 (0.08)	31.07 (0)	49.04 (0.12)	41.10 (0.02)	31.03 (0.04)	31.45 (0)	75.42 (1.26)	36.18 (0.28)
MLGPTQ (3-bit)	**31.72 (0.39)**	**26.93 (0.58)**	**30.11 (0.63)**	**27.97 (1.04)**	**30.77 (0.50)**	**28.06 (0.53)**	**47 (0.38)**	**39.07 (0.22)**	**27.62 (0.47)**	**28.74 (0)**	**54.84 (7.94)**	**31.90 (0.47)**
GPTQ$_{\text{tweaked}}$ (3-bit)	31.3 (0.62)	25.89 (0.7)	28.18 (0.75)	23.54 (1.01)	24.09 (0.51)	21.12 (0.39)	45.99 (0.26)	38.32 (0.15)	23.80 (0.46)	28.30 (0)	54.02 (7.9)	30.93 (0.24)
AWQ$_{\text{tweaked}}$ (3-bit)	31.57 (0.94)	26.45 (0.59)	25.13 (0.62)	24.46 (1.02)	26.77 (0.7)	19.79 (1.02)	45.13 (0.17)	37.62 (0.46)	21.83 (0.46)	28.24 (0)	53.66 (19)	31.03 (0.34)
RTN (3-bit)	26.02 (0.01)	0.03 (0)	0.03 (0)	0.06 (0)	0.05 (0)	0.05 (0)	0.9 (0)	0.10 (0)	0.34 (0)	26.38 (0)	51.3 (0)	31.4 (0)
GPTQ (3-bit)	30.83 (0.31)	26.19 (0.58)	31.88 (0.65)	28.21 (0.81)	32.93 (0.26)	29.75 (0.3)	47.22 (0.17)	39.53 (0.17)	26.12 (0.11)	28.16 (0)	56.03 (12.39)	31.26 (0.5)
AWQ (3-bit)	31.23 (0.63)	25.35 (0.18)	30.35 (0.73)	28.65 (1.15)	31.23 (0.32)	28.77 (0.62)	47.13 (0.17)	39.23 (0.44)	27.16 (0.57)	28.63 (0)	53.66 (4.08)	31.77 (0.29)