Table 1: Multilingual instruction following using Spanish Alpaca-GPT4 and English Alpaca-GPT4 data fine-tuning on Llama 3-8B. Evaluated with Arc challenge benchmark.

Models	English ↑	Spanish ↑
FedBCD	**59.73**	51.02
ParaBlock	59.47	**51.28**
Base	58.19	49.83

Table 2: Results under extreme data heterogeneity settings when fine-tuning Llama 3-8B on Alpaca-GPT4 and Math Instruct datasets.

Models	MT-Bench ↑	GSM8K ↑
i.i.d.	**5.21**	**56.14**
Non-i.i.d.	5.14	55.88
Extreme non-i.i.d.	5.13	54.66

Table 3: Ablation for block scheduling when fine-tuning Llama 3-8B on Alpaca-GPT4 and Math Instruct datasets.

Models	MT-Bench↑	GSM8K↑
Random	5.14	**55.88**
Seq.	5.03	54.21
Rev. seq.	5.06	54.59
Gradient-based	**5.19**	53.60

Table 4: Adaptation for extreme latency when fine-tuning Llama 3-8B on Math Instruct dataset.

Models	GSM8K↑	Runtime(m) ↓
ParaBlock (orignial)	**55.88**	38.72
ParaBlock (adapted)	52.46	**20.16**

Table 5: Results under cross-silo partial participation settings (50 clients 20% participation ratio) when fine-tuning Llama 3-8B on Alpaca-GPT4 and Math Instruct datasets.

Models	MT-Bench↑	GSM8K↑
Full participation	**5.14**	**55.88**
Cross-silo	5.10	53.90

Table 6: Integration with FedSparsify for fine-tuning Llama 3-8B on Math Instruct dataset.

Models	GSM8K↑	Runtime(m) ↓
ParaBlock	**55.88**	15.8
+FedSparsify-Global	Failed	Failed
+FedSparsify-Local	54.51	**15.7**

Table 7: Ablation for the number of layers in the block assignments when fine-tuning Llama 3-8B on Math Instruct dataset.

Models	GSM8K↑	Runtime(m) ↓
2 layers	**55.88**	15.8
Partial layer	53.22	15.0
1/4 layer	52.24	**14.9**