

## C More Experiments results

### C.1 Error-Bar of Current Experiments

This section presents error bar experiments, reporting the mean  $\pm$  standard error of the mean (SEM) over five independent runs, to substantiate the stability and performance of the FedASK framework. These evaluations cover varied differential privacy (DP) budgets and non-IID data distributions for Llama-2-7B and Llama-2-13B models on the dolly-15K and MetaMathQA datasets, respectively. Unless otherwise specified, all other parameters, such as batch size and communication rounds, align with the primary experimental configurations detailed in Section 5.

For these error-bar evaluations, specific learning rates and LoRA ranks were employed. Llama-2-7B experiments (Tables 5 and 7) used LoRA rank  $r = 64$ . For IID data with varying DP budgets (Table 5), baseline learning rates were  $2 \times 10^{-4}$  (non-DP) and  $1 \times 10^{-4}$  (DP); FedASK and other LoRA methods used  $5 \times 10^{-5}$  (non-DP) and  $4 \times 10^{-4}$  (DP) with gradient clipping of 1.0. For non-IID evaluations at DP  $\epsilon = 3$  (Table 7), baseline learning rates were  $1 \times 10^{-4}$ , and LoRA-based methods (including FedASK) used  $4 \times 10^{-4}$  with 1.0 gradient clipping. The Llama-2-13B experiments with IID data (Table 6) utilized a LoRA rank  $r = 128$ ; FedASK learning rates were  $5 \times 10^{-4}$  (non-DP) and  $4 \times 10^{-4}$  (DP), while baselines used  $2 \times 10^{-4}$ .

Table 5: Performance Comparison (Mean  $\pm$  SEM from five runs) on Llama-2-7B with Different DP Budgets.

Task	Priv. Budget	FedASK	FedAvg	FFA-LoRA	FedSA-LoRA	FedProx	Scaffold
MMLU	Non Private	46.34 $\pm$ 0.19	43.13 $\pm$ 2.01	45.72 $\pm$ 0.27	44.63 $\pm$ 0.57	43.55 $\pm$ 1.44	44.49 $\pm$ 1.17
	$\epsilon = 1$	45.79 $\pm$ 0.01	42.82 $\pm$ 0.75	44.15 $\pm$ 1.39	43.16 $\pm$ 0.25	43.00 $\pm$ 1.01	43.53 $\pm$ 0.12
	$\epsilon = 3$	45.88 $\pm$ 0.37	42.43 $\pm$ 0.94	43.82 $\pm$ 1.10	42.30 $\pm$ 1.17	43.07 $\pm$ 0.11	42.88 $\pm$ 0.41
	$\epsilon = 6$	45.73 $\pm$ 0.05	43.46 $\pm$ 0.12	44.02 $\pm$ 1.20	43.30 $\pm$ 0.54	43.00 $\pm$ 0.71	43.43 $\pm$ 0.37
DROP	Non Private	32.01 $\pm$ 0.08	30.31 $\pm$ 0.11	31.65 $\pm$ 0.31	30.86 $\pm$ 0.37	30.95 $\pm$ 0.04	30.87 $\pm$ 0.86
	$\epsilon = 1$	31.33 $\pm$ 0.10	30.49 $\pm$ 0.94	30.41 $\pm$ 1.31	29.96 $\pm$ 1.08	30.18 $\pm$ 0.67	30.08 $\pm$ 0.42
	$\epsilon = 3$	31.06 $\pm$ 1.02	29.69 $\pm$ 0.43	29.87 $\pm$ 1.47	29.66 $\pm$ 0.26	29.07 $\pm$ 0.57	28.90 $\pm$ 0.15
	$\epsilon = 6$	31.27 $\pm$ 0.10	29.46 $\pm$ 0.15	30.37 $\pm$ 0.97	29.91 $\pm$ 0.65	28.64 $\pm$ 1.07	30.19 $\pm$ 0.01
Human-Eval	Non Private	14.63 $\pm$ 0.61	13.42 $\pm$ 1.83	14.02 $\pm$ 0.02	12.81 $\pm$ 0.61	12.81 $\pm$ 0.61	14.63 $\pm$ 0.00
	$\epsilon = 1$	13.72 $\pm$ 1.52	11.28 $\pm$ 1.52	12.50 $\pm$ 0.30	11.28 $\pm$ 2.13	8.85 $\pm$ 3.35	8.54 $\pm$ 1.22
	$\epsilon = 3$	14.02 $\pm$ 1.22	7.63 $\pm$ 2.74	11.59 $\pm$ 0.61	8.85 $\pm$ 2.13	10.06 $\pm$ 3.35	9.76 $\pm$ 1.83
	$\epsilon = 6$	15.85 $\pm$ 0.02	9.76 $\pm$ 1.83	11.90 $\pm$ 0.31	10.67 $\pm$ 2.13	8.85 $\pm$ 2.13	9.76 $\pm$ 2.44

Table 6: Performance Comparison (Mean  $\pm$  SEM from five runs) on Llama-2-13B with Different DP Budgets.

Task	Priv. Budget	FedASK	FedAvg	FFA-LoRA	FedSA-LoRA	FedProx	Scaffold
GSM8K	Non-Private	51.40 $\pm$ 1.40	46.25 $\pm$ 2.25	48.50 $\pm$ 0.10	50.00 $\pm$ 2.80	47.95 $\pm$ 0.15	46.95 $\pm$ 1.35
	$\epsilon = 1$	24.95 $\pm$ 2.25	16.40 $\pm$ 0.90	14.25 $\pm$ 0.05	14.50 $\pm$ 2.30	16.00 $\pm$ 0.80	16.45 $\pm$ 0.35
	$\epsilon = 3$	25.35 $\pm$ 0.55	19.60 $\pm$ 3.10	18.60 $\pm$ 1.40	21.20 $\pm$ 1.00	20.20 $\pm$ 2.20	18.30 $\pm$ 2.50
	$\epsilon = 6$	24.35 $\pm$ 3.35	20.05 $\pm$ 0.75	19.15 $\pm$ 0.95	18.45 $\pm$ 1.15	22.05 $\pm$ 1.95	20.35 $\pm$ 0.05
GSM8K <sub>hard</sub>	Non-Private	23.90 $\pm$ 4.80	21.90 $\pm$ 3.90	22.20 $\pm$ 1.00	22.10 $\pm$ 1.30	23.05 $\pm$ 3.05	20.95 $\pm$ 0.85
	$\epsilon = 1$	13.65 $\pm$ 0.65	10.25 $\pm$ 1.45	7.85 $\pm$ 0.15	8.25 $\pm$ 1.65	7.90 $\pm$ 0.70	9.35 $\pm$ 0.25
	$\epsilon = 3$	13.00 $\pm$ 0.40	11.90 $\pm$ 0.80	10.25 $\pm$ 0.25	11.55 $\pm$ 0.25	11.35 $\pm$ 0.55	11.20 $\pm$ 0.20
	$\epsilon = 6$	13.30 $\pm$ 3.60	11.30 $\pm$ 0.80	9.40 $\pm$ 0.20	10.55 $\pm$ 0.35	11.60 $\pm$ 0.70	11.40 $\pm$ 0.60
Math	Non-Private	12.55 $\pm$ 0.75	9.30 $\pm$ 1.00	10.25 $\pm$ 0.55	10.60 $\pm$ 0.10	10.90 $\pm$ 0.80	10.00 $\pm$ 0.20
	$\epsilon = 1$	7.25 $\pm$ 0.35	5.55 $\pm$ 0.35	5.50 $\pm$ 0.30	5.85 $\pm$ 0.25	5.85 $\pm$ 0.25	5.55 $\pm$ 0.25
	$\epsilon = 3$	7.20 $\pm$ 0.60	6.50 $\pm$ 0.40	6.20 $\pm$ 0.20	6.15 $\pm$ 0.25	6.30 $\pm$ 0.10	5.95 $\pm$ 0.45
	$\epsilon = 6$	6.50 $\pm$ 1.10	6.35 $\pm$ 0.15	6.15 $\pm$ 0.15	6.05 $\pm$ 0.15	6.50 $\pm$ 0.20	7.00 $\pm$ 0.10

The inclusion of mean  $\pm$  SEM from five runs in these experiments offers robust statistical validation of the delineated advantages of FedASK, reinforcing the conclusions drawn from single-run experiments in the main paper. As detailed in Table 5 and Table 6, FedASK outperforms or performs comparable to baseline methods in non-private settings and DP budgets of  $\epsilon \in \{1, 3, 6\}$ , frequently producing comparable or reduced SEMs, highlighting its capacity to achieve an effective equilibrium between model utility and privacy preservation. This demonstrated robustness is further evident in scenarios characterized by data heterogeneity; Table 7 reveals FedASK’s consistent maintenance of leading average performance alongside constrained variability, as indicated by the SEMs, across diverse

610 non-IID Dirichlet distributions ( $\alpha \in \{0.1, 0.5, 1.0\}$ ), corroborating the adaptability observations  
611 presented in Section 5.2.

Table 7: Performance Comparison (Mean  $\pm$  SEM from five runs) for DP Budget  $\epsilon = 3$  across Different Data Distributions on Llama-2-7B.

Task	Data Dist.	FedASK	FedAvg	FFA-LoRA	FedSA-LoRA	FedProx	Scaffold
MMLU	IID	45.88 $\pm$ 0.37	42.43 $\pm$ 0.94	43.82 $\pm$ 1.10	42.30 $\pm$ 1.17	43.07 $\pm$ 0.11	42.88 $\pm$ 0.41
	Dir(0.1)	45.21 $\pm$ 0.84	42.85 $\pm$ 0.16	43.64 $\pm$ 1.09	44.11 $\pm$ 0.16	42.99 $\pm$ 0.39	42.33 $\pm$ 0.73
	Dir(0.5)	45.05 $\pm$ 0.90	42.71 $\pm$ 0.60	43.22 $\pm$ 1.76	43.26 $\pm$ 0.54	42.66 $\pm$ 0.32	42.91 $\pm$ 0.94
	Dir(1)	45.26 $\pm$ 0.75	42.75 $\pm$ 0.21	44.60 $\pm$ 1.37	42.37 $\pm$ 1.33	43.10 $\pm$ 0.12	42.69 $\pm$ 0.98
DROP	IID	31.10 $\pm$ 1.04	29.78 $\pm$ 0.34	29.87 $\pm$ 1.47	29.66 $\pm$ 0.26	29.07 $\pm$ 0.57	28.90 $\pm$ 0.15
	Dir(0.1)	30.85 $\pm$ 0.17	29.27 $\pm$ 0.93	30.52 $\pm$ 0.42	28.72 $\pm$ 0.14	28.53 $\pm$ 0.35	28.29 $\pm$ 0.02
	Dir(0.5)	31.15 $\pm$ 0.01	29.41 $\pm$ 0.23	29.98 $\pm$ 0.72	28.47 $\pm$ 0.64	28.47 $\pm$ 0.24	29.17 $\pm$ 0.28
	Dir(1)	31.19 $\pm$ 0.40	29.77 $\pm$ 0.25	30.16 $\pm$ 1.23	29.40 $\pm$ 0.11	29.85 $\pm$ 0.03	30.20 $\pm$ 0.32
Human-Eval	IID	14.02 $\pm$ 1.22	7.63 $\pm$ 2.74	11.59 $\pm$ 0.61	8.85 $\pm$ 2.13	10.06 $\pm$ 3.35	9.76 $\pm$ 1.83
	Dir(0.1)	13.41 $\pm$ 0.00	9.15 $\pm$ 1.83	11.29 $\pm$ 0.31	10.67 $\pm$ 2.13	9.15 $\pm$ 4.27	8.24 $\pm$ 0.31
	Dir(0.5)	12.81 $\pm$ 1.82	9.76 $\pm$ 1.22	11.89 $\pm$ 1.52	12.19 $\pm$ 0.61	14.33 $\pm$ 0.31	8.24 $\pm$ 1.52
	Dir(1)	13.41 $\pm$ 0.61	7.93 $\pm$ 1.83	12.50 $\pm$ 1.52	10.37 $\pm$ 0.00	7.93 $\pm$ 2.44	7.63 $\pm$ 1.52

## 612 C.2 Algorithms within more DP and Non-iid Conditions

613 Table 8 provides a comparative evaluation of algorithm performance when subjected to the combined  
614 effects of differential privacy and specific non-IID data distributions, namely Dirichlet distributions  
615 with  $\alpha = 0.1$  representing higher data heterogeneity and  $\alpha = 1.0$  indicating lower heterogeneity.  
616 This side-by-side presentation for each algorithm across various DP budgets ( $\epsilon \in \{1, 3, 6\}$  and  
617 Non-Private) allows for a nuanced understanding of their robustness.

618 The results in Table 8 underscore FedASK’s consistent ability to deliver strong performance even  
619 under these challenging compound conditions. Across the evaluated tasks (MMLU, BBH, DROP,  
620 Human-Eval), FedASK generally maintains a competitive edge or outperforms baseline methodolo-  
621 gies for both the more heterogeneous non-IID setting  $\alpha = 0.1$  and the less heterogeneous setting  
622  $\alpha = 1.0$ . In particular, while increased DP noise (smaller  $\epsilon$ ) or increased data heterogeneity (smaller  
623  $\alpha$ ) tends to degrade performance for all algorithms, FedASK often exhibits a more graceful degrada-  
624 tion compared to several baselines. This suggests that FedASK’s two-stage sketching and aggregation  
625 mechanism not only preserves utility under DP but also offers resilience against varying degrees of  
626 data heterogeneity. The comparative performance between the Non-IID 0.1 and Non-IID 1.0 columns  
627 for FedASK within each DP budget further illustrates its capacity to adapt effectively, reinforcing  
628 its suitability for practical federated learning scenarios where both privacy and non-IID data are  
629 prevalent concerns.

Table 8: Algorithm Performance across Varying DP Budgets for Non-IID (Dirichlet 0.1 and 1.0) Data on Llama-2-7B

Task	DP Setting	FedASK		FedAvg		FFA-LoRA		FedSA-LoRA		FedProx		Scaffold	
		$\alpha$ 0.1	$\alpha$ 1.0	$\alpha$ 0.1	$\alpha$ 1.0	$\alpha$ 0.1	$\alpha$ 1.0	$\alpha$ 0.1	$\alpha$ 1.0	$\alpha$ 0.1	$\alpha$ 1.0	$\alpha$ 0.1	$\alpha$ 1.0
MMLU	No DP	<b>46.50</b>	<b>46.32</b>	45.59	45.61	45.33	45.82	45.82	45.48	45.53	45.51	43.75	44.70
	DP $\epsilon = 1$	<b>45.73</b>	<b>45.88</b>	42.69	42.12	41.00	43.75	41.44	41.62	41.40	43.66	43.11	43.60
	DP $\epsilon = 3$	<b>46.04</b>	<b>45.86</b>	42.69	42.96	42.54	43.23	44.27	41.04	42.61	42.98	43.05	41.71
	DP $\epsilon = 6$	<b>46.24</b>	<b>46.42</b>	41.79	43.97	42.82	42.24	42.94	43.93	40.07	43.56	41.06	43.83
BBH	No DP	<b>32.15</b>	<b>32.55</b>	33.50	32.03	32.16	32.86	32.25	32.11	33.10	32.29	32.99	33.08
	DP $\epsilon = 1$	<b>31.99</b>	<b>31.78</b>	31.14	31.73	31.71	33.45	33.34	31.80	32.22	33.40	32.28	31.36
	DP $\epsilon = 3$	<b>32.46</b>	<b>32.39</b>	33.54	31.02	32.71	32.06	33.37	32.61	32.30	32.17	33.02	31.50
	DP $\epsilon = 6$	<b>32.12</b>	<b>31.91</b>	31.98	30.99	32.42	31.56	31.00	31.65	31.67	34.01	31.12	32.44
DROP	No DP	<b>33.16</b>	<b>32.30</b>	30.56	31.47	31.19	33.46	30.96	31.92	31.09	32.78	28.94	30.12
	DP $\epsilon = 1$	<b>31.93</b>	<b>31.98</b>	28.18	29.93	29.75	31.17	28.96	28.81	28.16	30.71	29.45	29.20
	DP $\epsilon = 3$	<b>31.01</b>	<b>31.09</b>	31.49	30.02	30.10	28.93	28.58	29.29	28.18	29.82	28.27	30.52
	DP $\epsilon = 6$	<b>32.46</b>	<b>30.90</b>	28.78	28.35	29.51	29.21	30.37	26.51	29.96	28.51	28.88	29.80
Human-Eval	No DP	<b>15.24</b>	<b>15.85</b>	12.20	12.80	14.63	14.02	14.02	14.02	12.80	14.02	12.80	15.85
	DP $\epsilon = 1$	<b>14.02</b>	<b>12.20</b>	11.59	10.98	9.76	12.20	10.98	10.98	12.20	12.80	14.02	12.80
	DP $\epsilon = 3$	<b>12.20</b>	<b>13.41</b>	12.20	9.76	10.98	10.98	12.80	10.37	13.41	10.37	7.93	9.15
	DP $\epsilon = 6$	<b>12.20</b>	<b>13.41</b>	6.71	13.41	10.37	10.37	6.71	10.98	11.59	12.80	14.63	12.80

### 630 C.3 Sensitive Experiments

#### 631 C.3.1 Choice on the lora rank

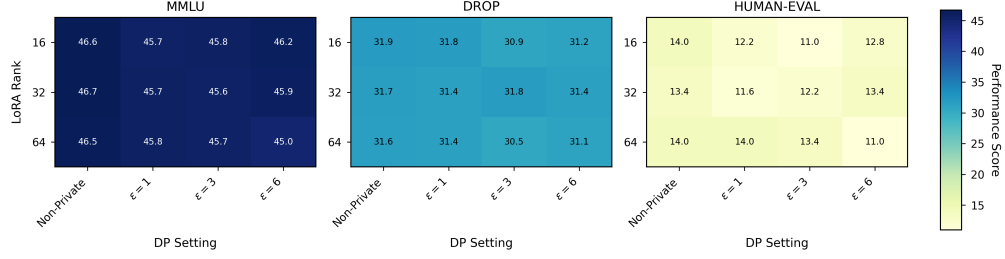


Figure 4: Performance of Llama 2-7B on IID data across LoRA ranks and differential privacy (DP) settings ( $\epsilon$ ) for MMLU, DROP, and Human tasks.

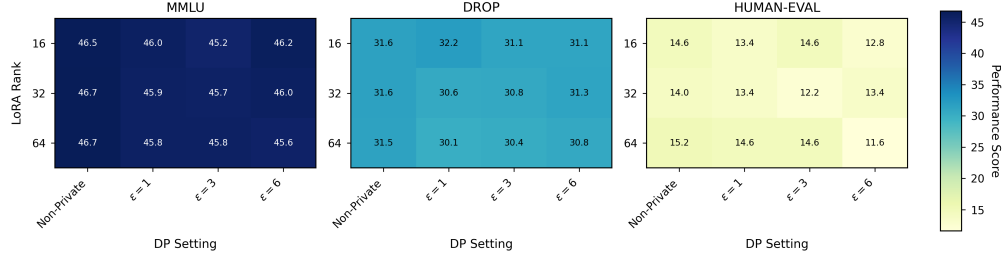


Figure 5: Performance of Llama 2-7B on Non-IID data ( $\alpha = 0.1$ ) across LoRA ranks and differential privacy (DP) settings ( $\epsilon$ ) for MMLU, DROP, and Human tasks.

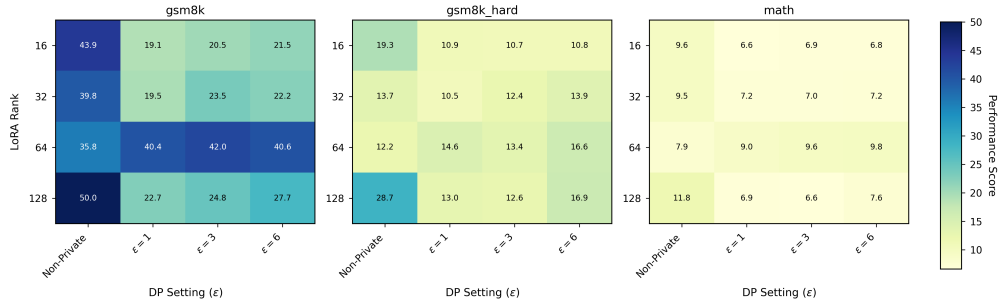


Figure 6: Performance of Llama 2-13B on IID data across LoRA ranks and differential privacy (DP) settings ( $\epsilon$ ) for gsm8k, gsm8k\_hard, and math tasks.

632 The selection of an appropriate LoRA rank  $r$  is crucial to balance the performance of the model  
 633 and the efficiency of the parameters, particularly when differential privacy (DP) is applied. This  
 634 section details the interaction between LoRA rank, DP settings, model size, and data distribution for  
 635 the FedASK framework, referencing empirical results from Llama 2-7B and Llama-2-13B models.  
 636 Although higher LoRA ranks, such as  $r = 128$ , often deliver superior performance in nonprivate  
 637 scenarios, the introduction of DP mechanisms significantly alters these performance landscapes, often  
 638 favoring intermediate ranks for a more robust utility-privacy trade-off.

639 A key finding, illustrated in Figure 6 for the Llama 2-13B model on IID data, is the change in the  
 640 optimal LoRA rank under DP for FedASK. Although rank 128 excels without privacy, for instance, on  
 641 gsm8k (50.0) and gsm8k\_hard (28.7), intermediate ranks frequently provide a superior utility-privacy  
 642 trade-off when DP is enabled. Specifically, on the gsm8k task, rank 64 consistently outperforms rank

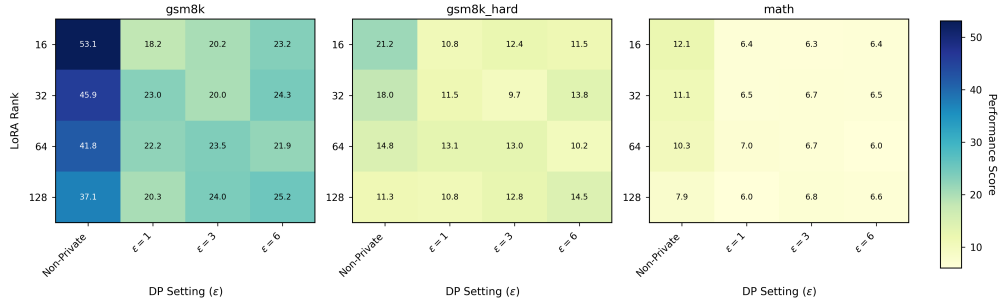


Figure 7: Performance of Llama 2-13B on Non-IID data ( $\alpha = 0.1$ ) across LoRA ranks and differential privacy (DP) settings ( $\epsilon$ ) for gsm8k, gsm8k\_hard, and math tasks.

128 under all tested DP settings; for example, with  $\epsilon = 1$ , rank 64 achieves 40.4 versus 22.7 for rank 128, and with  $\epsilon = 6$ , rank 64 achieves 40.6 versus 27.7 for rank 128. Similarly, for the math task, rank 64 shows better performance than rank 128 across all DP budgets. On gsm8k\_hard, rank 64 also remains highly competitive with, or slightly better than, rank 128 under DP conditions. This consistent strong performance of rank 64 under various DP constraints suggests that for FedASK with the 13B model on IID data, a moderately sized LoRA rank can be more parameter-efficient and achieve better utility when stringent privacy guarantees are necessary. When data heterogeneity is introduced for the Llama 2-13B model, as shown in Figure 7 for Non-IID data ( $\alpha = 0.1$ ), the utility of intermediate ranks under DP persists largely. Although overall performance levels may adjust due to the non-IID distribution, FedASK with moderately sized ranks continues to demonstrate a strong balance between adaptation capability and resilience to DP noise, reinforcing the notion that maximal ranks are not universally optimal under privacy constraints in heterogeneous settings.

This rank-dependent performance pattern under DP is also investigated for the Llama 2-7B model. Figure 4 presents results on IID data for MMLU, DROP, and HumanEval tasks. For FedASK, it is generally observed that while larger ranks might offer marginal gains or lead in non-private scenarios, the application of DP tends to make intermediate ranks more advantageous. These moderately sized ranks appear to strike an effective balance, providing sufficient capacity for task adaptation while mitigating the detrimental impact of DP noise that can be more pronounced with a larger number of trainable parameters. The introduction of significant data heterogeneity with Non-IID data ( $\alpha = 0.1$ ), illustrated in Figure 5, further tests this dynamic. Even in these challenging conditions, FedASK with intermediate ranks often maintains robust performance relative to larger ranks under DP. This suggests that for the 7B model, an excessively large rank under combined DP and non-IID stress may not yield proportional benefits and could be outperformed by more parameter-efficient intermediate rank configurations.

### C.3.2 Choice on over-sketching rate

The precision of FedASK’s aggregation mechanism is theoretically linked to the choice of the over-sketching parameter  $p$ , which, together with the LoRA rank  $r$ , defines the sketching dimension  $r + p$ . While Theorem 2 provides a condition for exact aggregation, it is crucial to empirically assess the impact of varying sketching dimensions on aggregation fidelity under practical conditions, including different degrees of data heterogeneity and client participation numbers. This appendix section details these specific evaluations for FedASK, illustrating its robustness. The experiments summarized here were conducted to determine a suitable range for the sketching dimension, ensuring high fidelity without unnecessary computational overhead. All results presented pertain to the FedASK algorithm, and aggregation fidelity is quantified as the cosine similarity between the global LoRA update reconstructed by FedASK and the ideal average of true local LoRA updates.

The empirical investigation involved evaluating the aggregation fidelity of FedASK across a matrix of conditions, as depicted in Figure 9. The experiments systematically varied: (i) the sketching dimension (x-axis values: 6, 32, 51, 64, and 96), (ii) the degree of non-iid degree (Dirichlet distributions with  $\alpha \in \{1.0, 0.8, 0.5, 0.1\}$ ), and (iii) the number of participating clients, shown in four distinct panels: (a) 5 clients, (b) 10 clients, (c) 15 clients, and (d) 20 clients. For these specific

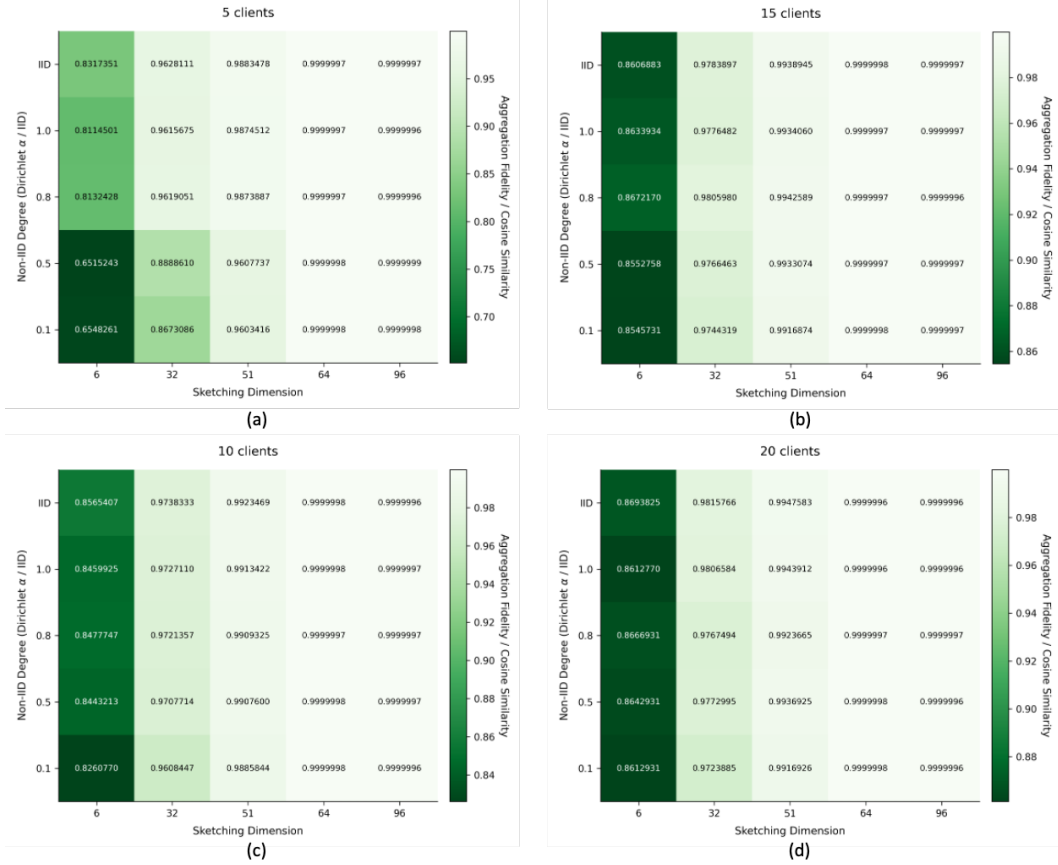


Figure 8: Impact of sketching dimension (x-axis) and non-IID degree (y-axis, Dirichlet  $\alpha$  / IID) on FedASK’s aggregation fidelity (cosine similarity) for (a) 5, (b) 10, (c) 15, and (d) 20 clients, showing robust near-unity performance.

fidelity evaluations, differential privacy mechanisms were not applied to isolate the performance of the aggregation mechanism itself. The color intensity in each heatmap cell corresponds to the achieved cosine similarity, with lighter shades indicating higher fidelity.

The results consistently demonstrate FedASK’s exceptional aggregation fidelity across the vast majority of tested scenarios. As seen in Figure 8, near-unity cosine similarity is achieved for most combinations of sketching dimensions, non-IID degrees, and client numbers. Even with the smallest sketching dimensions, fidelity remains remarkably high, particularly as the number of participating clients increases (panels b, c, and d). While the 5-client scenario (panel a) shows slightly reduced fidelity under extreme non-IID conditions and very small sketching dimensions, the performance rapidly approaches unity with modest increases in either parameter. These findings underscore that FedASK is not highly sensitive to the over-sketching rate for maintaining precise aggregation and can achieve excellent fidelity even with minimal or conservative sketching dimensions, confirming its practical efficiency and robustness.

#### C.4 System Efficiency Experiments

To assess the system efficiency of the federated learning algorithms evaluated, we measured key resource utilization metrics for a single client operating within a federated network of 5 clients. These experiments were carried out on NVIDIA H100 GPUs. The primary metrics, communication volume and peak GPU memory consumption, are detailed for Llama 2-7B and Llama 2-13B models trained with 4-bit precision.

Figure 9 illustrates the usage of resources per client. The volume of communication, segmented into uplink and downlink traffic, is reported in millions of parameters. Peak

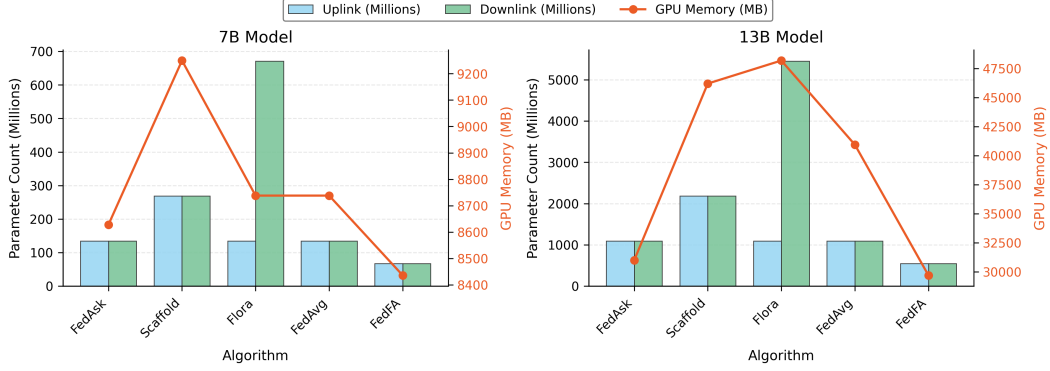


Figure 9: System resource utilization for five federated learning algorithms when training Llama 2-7B (left) and Llama 2-13B (right) models using 4-bit precision. The metrics, shown for a single client in a 5-client federated setup, include uplink and downlink communication volume (millions of parameters) and GPU memory consumption (MB).

Table 9: Total communication volume per client round for federated learning algorithms using Llama 2-7B and Llama 2-13B models under different numerical precisions. All values are in Megabytes (MB).

Algorithm	Llama 2-7B (MB)			Llama 2-13B (MB)		
	FP16	INT8	INT4	FP16	INT8	INT4
FedAsk	512	256	128	4160	2080	1040
Scaffold	1024	512	256	8320	4160	2080
Flora	1536	768	384	12480	6240	3120
FedAvg	512	256	128	4160	2080	1040
FedFA	256	128	64	2080	1040	520

GPU memory consumption (in MB) was meticulously monitored on the client side using the `torch.cuda.max_memory_allocated(device=torch.device('cuda'))` PyTorch function, capturing the maximum memory footprint during local training operations with differential privacy mechanisms enabled. Furthermore, Table 9 quantifies the total communication volume (uplink plus downlink, in MB) per client round with different numerical precisions (FP16, INT8, and INT4) for both models Llama 2 sizes. This provides a comparative view of how the precision of the data impacts the communication overhead for each algorithm.

System efficiency evaluations highlight distinct resource profiles. FedFA achieves the lowest communication parameter counts by freezing its A matrix, as shown in Figure 9. FedASK, engineered to update both LoRA adaptor matrices for enhanced learnability, offers substantial communication efficiency; its INT4 communication volume detailed in Table 9 is approximately 50% less than Scaffold and roughly 66.7% less than Flora. Regarding GPU memory, Figure 9 indicates FedASK’s footprint is comparable to FedAvg and Scaffold but notably less than Flora—for instance, around 23.9% less for the Llama 2-13B model.