1 Appendix

1.1 Additional Experimental Setup

We highlight some important hyperparamters in fine-tuning the downstream tasks. For the Named Entity Recognition (NER) task on the JNLPBA [Collier and Kim, 2004] dataset in the Biology domain, we set the seed to 40, batch size to 32, and learning rate to 4×10^{-5} for in total of 15 epochs. For the same task on the SciERC [Luan *et al.*, 2018] dataset in the Computer Science domain, we set the seed to 40, batch size to 2, and learning rate to 7×10^{-5} for 15 epochs. On the other hand, for the classification (CLS) task on the Rct-20k [Beltagy *et al.*, 2019] dataset in the Medicine domain, we tune the seed to 40, batch size to 2, and learning rate to 4×10^{-6} for 15 epochs.

1.2 Additional Results for Ablation Study

In supplement to the ablation study in Section 5.3, we report the results on the JNLPBA dataset in the Biology domain in Table 1 and those on the Rct-20k dataset in the Medicine domain in Table 2.

Models	Computational Cost (unit: second)	Communication Cost (unit: MB)	Accuracy
FedBFPT-ALL	786.86(63.59%)	255.57(61.17%)	0.7196
FedBFPT\PL	**577.85(46.70%)**	**29.42(7.04%)**	0.7145
FedBFPT\M	**577.85(46.70%)**	**29.42(7.04%)**	0.7151
FedBFPT\S	**577.85(46.70%)**	**29.42(7.04%)**	0.7169
FedBFPT (ours)	**577.85(46.70%)**	**29.42(7.04%)**	**0.7198**

Table 1: Ablation Study (Biology)

Models	Computational Cost (unit: second)	Communication Cost (unit: MB)	Accuracy
FedBFPT-ALL	786.86(63.59%)	255.57(61.17%)	0.8185
FedBFPT\PL	**577.85(46.70%)**	**29.42(7.04%)**	0.8073
FedBFPT\M	**577.85(46.70%)**	**29.42(7.04%)**	0.8153
FedBFPT\S	**577.85(46.70%)**	**29.42(7.04%)**	0.8169
FedBFPT (ours)	**577.85(46.70%)**	**29.42(7.04%)**	**0.8312**

Table 2: Ablation Study (Medicine)

The results on both the Biology and Medicine datasets demonstrate that our proposed FedBFPT achieves the least computational and communication cost as well as the highest accuracy. These results are consistent with what we have presented in Section 5.3.

1.3 Additional Results for Parameter Study

Effect of the Number of Trained T-layers. Varying the number of the trained layers from 4 to 8, the measures on the Biology and Medicine datasets are reported in Tables 3 and 4, respectively. The additional results exhibit a similar trend as that reported in Section 5.4, indicating that our chosen FedBFPT-6 achieves a good balance between the overhead and the accuracy.

Models	Computational Cost (unit: second)	Communication Cost (unit: MB)	Accuracy
FedBFPT-4	**463.87(37.49%)**	**29.42(7.04%)**	0.7140
FedBFPT-8	693.62(56.05%)	**29.42(7.04%)**	0.7147
FedBFPT-6 (ours)	577.85(46.70%)	**29.42(7.04%)**	**0.7198**

Table 3: Effect of Varying the Number of Trained T-layers (Biology)

Models	Computational Cost (unit: second)	Communication Cost (unit: MB)	Accuracy
FedBFPT-4	**463.87(37.49%)**	**29.42(7.04%)**	0.8137
FedBFPT-8	693.62(56.05%)	**29.42(7.04%)**	0.8248
FedBFPT-6 (ours)	577.85(46.70%)	**29.42(7.04%)**	**0.8312**

Table 4: Effect of Varying the Number of Trained T-layers (Medicine)

Effect of Skewed Local Datasets. Corresponding to the results reported in Table 6 in Section 5.4, the sizes of uniform datasets and skewed datasets are shown in Table 5. In skewed datasets, there are large differences in size between the local datasets at different clients.

Client	Uniform(unit: MB)			Skewed(unit: MB)		
	Biology	CS	Medicine	Biology	CS	Medicine
Client-1	72.59	59.35	149.36	20.63	17.25	43.70
Client-2	72.64	60.69	151.47	66.71	56.11	141.27
Client-3	72.76	61.44	152.82	128.37	108.24	272.47
Client-4	71.61	61.89	151.78	128.42	108.23	272.21
Client-5	69.86	58.19	153.92	66.76	56.11	141.25
Client-6	72.01	61.75	149.69	20.59	17.37	43.62
Center	431.48	363.30	914.45	431.48	363.30	914.45

Table 5: Sizes of Uniform and Skewed Datasets

References

[Beltagy *et al.*, 2019] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[Collier and Kim, 2004] Nigel Collier and Jin-Dong Kim. Introduction to the bio-entity recognition task at JNLPBA. In *NLPBA/BioNLP*, pages 73–78, 2004.

[Luan *et al.*, 2018] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*, 2018.