Methods	Restaurant		Laptop		Twitter	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**
DeepSeek-1.5B	68.5	44.4	64.4	44.8	42.8	33.5
DeepSeek-7.0B	74.5	50.2	73.6	50.1	47.7	35.9
Ours	**87.22**	**82.08**	**82.28**	**78.83**	**77.84**	**76.85**

To ensure a thorough evaluation of our method, we also evaluate the performance of **deepseek-1.5b** and **-7.0b** on the benchmarks. The two LLMs are locally deployed in our server using Ollama, and the results are summarized in Tab. In general, GFN-ASSG significantly outperforms the DeepSeek models in terms of both accuracy and F1. On the Twitter dataset, our method has over 30% improvements in F1 than **DeepSeek-7.0b**, indicating that task-specific optimization is still required even for LLM.