A Detailed Result for Each Dataset

Table 1 and Table 2 report the main performance of few-shot supervised learning and few-shot in-context learning across three different models under label perturbations per dataset.

Dataset	**LM**	**Noisy Level**									
		0		25		50		75		100	
		SL	ICL	SL	ICL	SL	ICL	SL	ICL	SL	ICL
MR	GPT2-Large	53.4	64.7	45.3	51.4	48.1	47.9	45.5	44.2	38.2	41.1
	GPT2-XL	52.1	46.0	44.0	52.0	47.2	50.5	44.2	47.6	38.4	50.1
	GPT-J	48.5	54.6	44.8	36.9	48.7	40.4	40.8	40.7	34.2	39.7
SST2	GPT2-Large	52.1	62.2	44.1	61.1	43.4	45.0	44.4	39.0	46.8	42.9
	GPT2-XL	48.7	55.2	47.6	49.3	47.2	41.8	45.6	37.4	42.4	42.3
	GPT-J	50.2	67.1	44.2	63.3	45.8	47.5	43.9	38.9	38.9	36.4
RTE	GPT2-Large	49.3	34.9	46.9	38.8	49.2	40.5	47.4	39.4	47.4	39.4
	GPT2-XL	47.3	37.8	43.5	39.9	45.5	37.7	47.3	37.8	48.1	37.8
	GPT-J	43.0	40.3	42.5	43.8	42.1	42.6	41.0	36.1	41.9	37.3
CB	GPT2-Large	47.0	19.9	32.1	18.9	27.2	17.5	22.2	15.9	17.2	17.3
	GPT2-XL	47.9	22.7	35.6	26.7	26.4	24.3	21.4	22.5	15.6	22.8
	GPT-J	45.7	31.5	41.7	27.6	33.1	28.9	23.6	26.6	17.1	24.7
AG-NEWS	GPT2-Large	54.7	34.8	41.7	37.3	26.6	32.7	20.0	33.1	14.2	28.9
	GPT2-XL	51.9	37.6	39.9	42.6	25.9	31.8	22.9	30.9	13.5	28.9
	GPT-J	56.8	53.3	50.7	47.4	30.9	32.8	20.1	27.8	9.9	26.2
TREC	GPT2-Large	26.2	10.9	27.1	10.8	21.1	8.2	13.4	7.3	9.9	6.9
	GPT2-XL	25.8	40.8	24.2	33.1	21.2	30.9	16.7	24.9	10.0	19.2
	GPT-J	24.8	36.9	23.2	33.1	20.2	25.8	13.7	15.4	10.3	11.9

Table 1: Performance comparison of supervised learning and in-context learning on noisy labels across all six text classification datasets.We show the Macro-F1 score.

Dataset	LM	Imbalance Ratio					
		Low		Medium		High	
		SL	ICL	SL	ICL	SL	ICL
MR	GPT2-Large	58.9	59.1	55.2	60.0	36.2	58.4
	GPT2-XL	59.2	46.3	49.9	58.3	34.4	44.4
	GPT-J	58.8	59.4	41.1	54.3	33.8	51.3
GLUE-SST2	GPT2-Large	53.3	51.5	44.0	50.0	36.8	45.3
	GPT2-XL	48.8	42.4	47.9	52.3	37.0	37.9
	GPT-J	51.8	59.8	43.4	70.3	34.4	74.8
GLUE-RTE	GPT2-Large	46.2	36.8	44.1	38.3	36.3	34.5
	GPT2-XL	47.2	34.6	41.4	47.1	34.5	35.6
	GPT-J	46.6	38.4	44.4	36.3	34.8	34.5
SuperGLUE-CB	GPT2-Large	47.7	22.2	48.6	15.6	45.3	21.5
	GPT2-XL	45.2	31.8	47.4	27.5	48.1	23.8
	GPT-J	48.4	31.6	48.4	33.3	46.4	28.6
AG-NEWS	GPT2-Large	58.2	42.4	56.2	41.3	36.2	33.1
	GPT2-XL	49.9	44.6	59.6	31.0	46.0	40.2
	GPT-J	58.5	66.7	63.0	67.6	47.6	51.9
TREC	GPT2-Large	38.5	12.1	26.7	12.4	18.0	10.9
	GPT2-XL	38.7	42.4	34.0	48.8	21.4	45.2
	GPT-J	35.4	47.9	32.2	54.6	14.9	34.1

Table 2: Performance comparison of supervised learning and in-context learning on label imbalance across all six text classification datasets.We show the Macro-F1 score.

B Visualization of Performance Comparison of SL and ICL with Different Classification Types

Figure 1 and Figure 2 provide the visualization of the performance comparison of SL and ICL with different classification types under noisy labels and imbalanced ratios respectively.

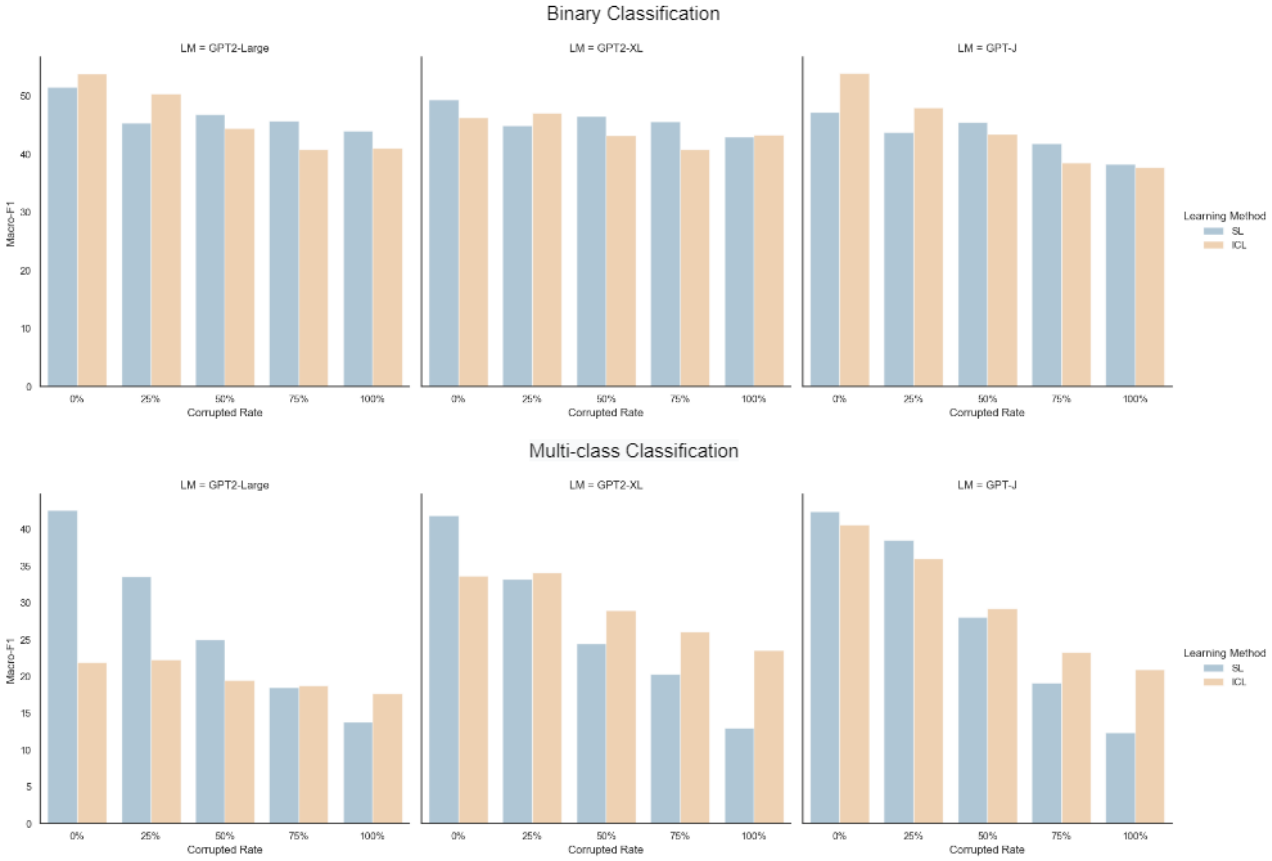

Figure 1: Performance comparison of supervised learning and in-context learning with different classification types under different noisy levels.

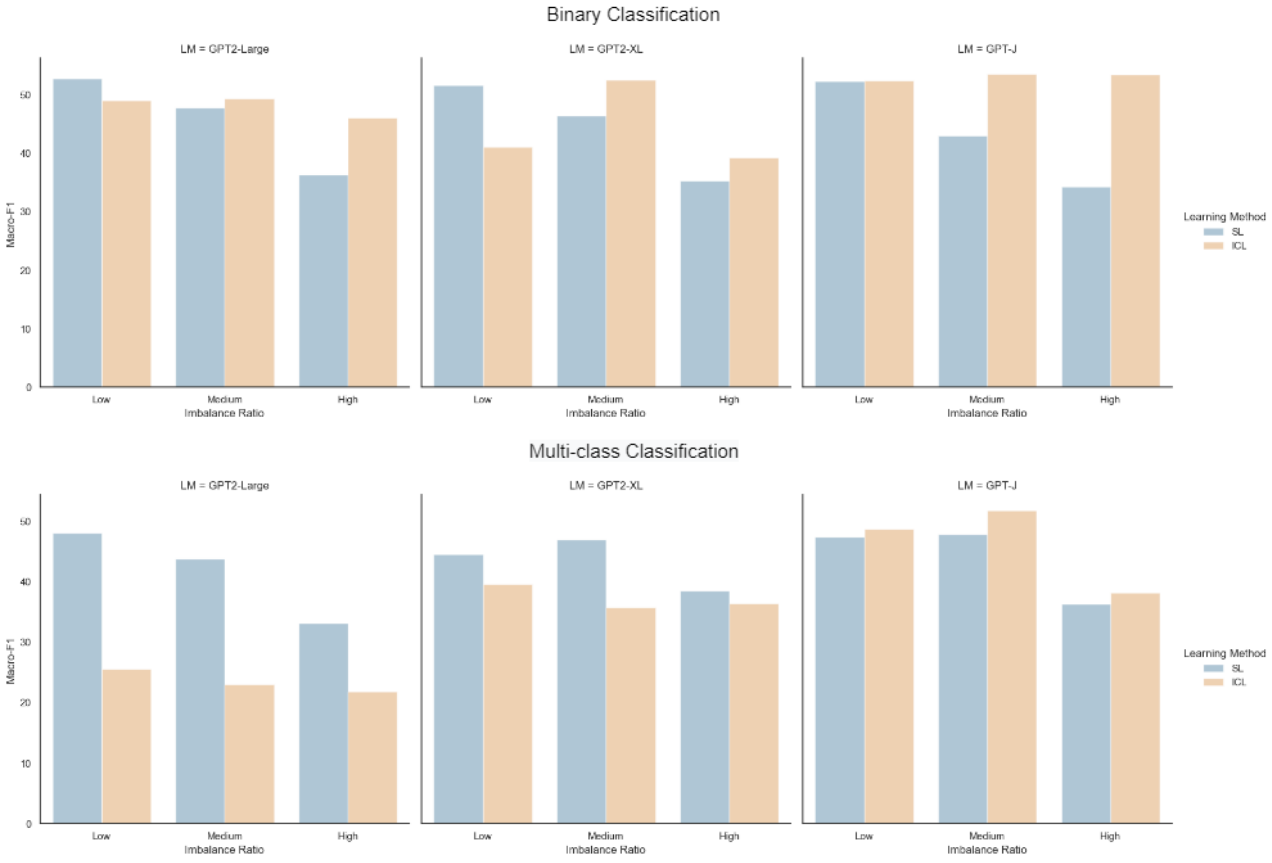

Figure 2: Performance comparison of supervised learning and in-context learning with different classification types under different imbalance ratios.