Table 1: Overview of CoLLaM in Comparison with Existing General and Legal Domain Benchmarks.

Dataset	Language	Domain	Data Source	Taxonomy	Task Num	Data Size
C-Eval_Legal	Chinese	General	existing datasets	-	2	493
CMMLU_Legal	Chinese	General	existing datasets	-	1	216
LEAGALBENCH	English	Legal	existing datasets expert annotation	issue-spotting rule-recall rule-application rule-conclusion	162	-
LawBench	Chinese	Legal	existing datasets	Memorization Understanding Applying	20	10000
LAiW	Chinese	Legal	existing datasets	basic information retrieval legal foundation inference complex legal application	14	11605
CoLLaM	Chinese	Legal	existing datasets, exam, expert annotation	Memotization Understanding Logic Inference Discrimination Generation Ethic	23	13650

Table 2: Evaluation Results of Generation Tasks Across Six LLMs with Human Annotations. For convenience, we attach the average scores and ranks for BERTScore and BARTScore.

Model	5-1	5-2	5-3	5-4	Human/Rank	BertScore/Rank	BartScore/Rank	Rough-L/Rank
GPT-4	4.12	3.58	4.26	3.24	3.80/1	0.774/1	-4.103/1	0.291/4
Qwen-14B-Chat	4.06	3.72	4.2	3.08	3.77/2	0.768/2	-4.226/3	0.328/2
Qwen-7B-Chat	4.04	3.54	4.16	3.10	3.71/3	0.758/5	-4.337/5	0.323/3
ChatGPT	4.08	3.40	4.04	3.24	3.69/4	0.759/3	-4.230/4	0.275/5
Fuzi-Mingcha	4.48	3.40	3.26	2.90	3.51/5	0.759/4	-4.192/2	0.350/1
ChatLaw-33B	3.24	1.56	3.64	2.38	2.71/6	0.680/6	-6.202/6	0.189/6

Table 3: Performance Comparison of Fine-tuned BERT and GPT-4 on CoLLaM.

Model	Memorization(Acc.)			Understanding(Acc.)					Logic Inference(Acc.)					
	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	3-1	3-2	3-3	3-4	3-5	3-6
BERT	17.2	9.3	10.0	9.0	13.3	13.0	22.2	20.2	19.1	21.1	4.8	18.8	26.1	16.8
GPT-4	27.2	34.8	14.0	79.8	51.0	94.0	77.2	96.2	79.2	68.3	62.4	33.2	66.0	51.0

Table 4: Performance Comparison of Fine-tuned BERT and GPT-4 on CoLLaM.

Model	Discrimination(Acc.)		Ethic(Acc.)			Average
	4-1	4-2	6-1	6-2	6-3	
BERT	45.0	17.1	9.1	22.2	27.2	18.0
GPT-4	34.0	39.1	65.2	55.2	75.8	58.8

Table 5: Zero-shot performance(%) of various models at Memorization, Understanding, and Logic Inference level. Mirco-F1 was reported to provide a more comprehensive evaluation.

Model	Memorization(F1)			Understanding(F1)					Logic Inference(F1)					
	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	3-1	3-2	3-3	3-4	3-5	3-6
GPT-4	53.1	34.8	51.0	86.0	60.2	95.5	86.2	96.2	48.1	75.5	62.4	55.8	66.2	78.7
Qwen-14B-Chat	57.1	38.6	46.0	93.8	51.9	90.7	86.4	92.1	90.3	90.5	66.4	53.3	44.8	49.2
Qwen-7B-Chat	49.4	38.8	40.6	80.5	51.8	88.0	67.2	92.7	89.7	84.0	25.8	47.0	36.7	31.7
ChatGPT	45.0	25.6	38.2	73.9	49.1	87.0	76.2	82.2	87.9	60.3	47.6	38.2	39.6	43.0
Biachuan-13B-Chat	37.4	33.9	40.3	59.1	41.2	72.0	63.5	75.4	87.0	57.8	41.8	38.4	33.5	21.5
InternLM-7B-Chat	47.1	36.7	40.5	77.1	47.6	88.0	68.9	61.2	85.2	76.6	22.8	45.7	37.3	42.6
ChatGLM3	45.3	28.7	36.2	56.8	41.4	69.0	64.0	79.4	81.3	58.9	17.0	40.4	24.9	37.6
ChatGLM2	59.1	26.5	54.8	58.8	40.0	67.4	53.0	67.1	88.3	58.3	3.8	50.6	30.3	34.9
Baichuan-13B-base	55.2	24.8	45.7	58.2	37.8	77.2	59.2	74.4	72.6	42.4	57.9	47.7	31.1	20.4
Chinese-Alpaca-2-7B	45.7	28.8	52.1	59.6	36.8	64.0	54.7	30.9	71.7	48.5	60.2	41.3	22.0	15.5
Fuzi-Mingcha	43.7	30.3	41.1	67.5	35.9	67.8	48.7	36.0	80.0	58.8	15.7	41.7	29.8	29.3
ChatLaw-33B	41.2	24.3	36.4	57.8	34.5	77.0	62.9	75.5	77.6	42.1	3.5	37.1	28.6	30.3
InternLM-7B	55.5	40.0	46.3	43.0	40.8	60.4	64.6	58.4	81.7	44.7	65.5	48.4	35.0	22.0
TigerBot-Base	46.4	28.4	43.5	52.6	37.0	61.0	59.2	24.4	80.9	36.8	26.2	46.0	30.7	26.0

Table 6: Zero-shot performance(%) of various models at Memorization, Understanding, and Logic Inference level. Mirco-F1, BERTScore, BARTScore are reported to provide a more comprehensive evaluation.

Model	Discrimination(F1)		Generation(BERTScore)				Generation(BARTScore)				Ethic(F1)		
	4-1	4-2	5-1	5-2	5-3	5-4	5-1	5-2	5-3	5-4	6-1	6-2	6-3
GPT-4	31.3	72.0	0.711	0.713	0.860	0.679	-4.18	-5.01	-5.60	-3.93	82.6	77.2	88.3
Qwen-14B-Chat	28.7	57.0	0.747	0.735	0.847	0.697	-3.66	-4.93	-5.83	-4.16	66.1	65.9	78.1
Qwen-7B-Chat	29.5	54.9	0.740	0.721	0.828	0.694	-3.63	-5.51	-6.31	-4.19	46.1	61.8	73.3
ChatGPT	30.0	46.1	0.687	0.583	0.836	0.674	-4.40	-6.48	-5.74	-3.80	60.3	59.5	72.8
Biachuan-13B-Chat	27.4	43.6	0.739	0.637	0.843	0.661	-4.20	-13.47	-6.12	-5.05	44.6	47.6	58.8
InternLM-7B-Chat	0.2	34.8	0.588	0.342	0.794	0.616	-4.66	-11.43	-6.74	-4.90	53.7	57.4	71.1
ChatGLM3	25.7	40.0	0.729	0.713	0.806	0.674	-3.81	-4.93	-6.71	-3.99	56.7	56.6	69.6
ChatGLM2	22.7	46.4	0.736	0.696	0.774	0.672	-3.62	-4.87	-7.11	-3.91	67.2	60.3	73.1
Baichuan-13B-base	18.1	48.7	0.678	0.641	0.738	0.686	-4.31	-13.15	-6.58	-5.26	52.8	56.8	68.0
Chinese-Alpaca-2-7B	26.5	46.4	0.727	0.699	0.822	0.656	-3.79	-5.19	-6.54	-4.07	53.9	47.9	52.2
Fuzi-Mingcha	19.5	50.8	0.809	0.672	0.681	0.679	-2.90	-7.03	-7.12	-4.56	39.1	46.2	45.3
ChatLaw-33B	8.2	43.5	0.658	0.610	0.709	0.633	-4.41	-10.45	-7.68	-4.82	42.4	48.7	50.3
InternLM-7B	6.4	49.9	0.495	0.567	0.693	0.587	-5.32	-8.03	-7.62	-4.85	65.9	62.6	73.8
TigerBot-Base	27.3	52.1	0.673	0.495	0.787	0.658	-4.41	-6.65	-6.82	-4.05	61.4	51.8	60.8

Table 7: Zero-shot performance(%) of various models at Memorization, Understanding, and Logic Inference level. Mirco-F1 was reported to provide a more comprehensive evaluation.

Model	Memorization(F1)			Understanding(F1)					Logic Inference(F1)					
	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	3-1	3-2	3-3	3-4	3-5	3-6
ChatGLM	38.7	30.2	39.6	56.0	39.9	75.1	45.2	35.1	69.6	53.6	27.3	39.7	30.6	27.1
XVERSE-13B	60.8	38.7	60.4	45.6	41.1	70.0	46.3	30.2	64.1	45.0	36.4	53.6	29.8	21.9
Chinese-LLaMA-2-7B	39.8	27.3	41.8	64.5	25.3	60.0	31.2	24.2	64.3	42.9	41.7	34.5	29.7	1.6
Chinese-LLaMA-2-13B	40.6	5.1	28.6	39.7	33.7	79.0	61.4	41.2	64.3	38.7	0.0	35.0	19.3	31.1
Ziya-LLaMA-13B	40.0	27.0	40.5	75.6	39.1	56.5	53.4	25.5	74.2	46.5	4.5	33.4	30.4	26.4
LexiLaw	37.9	27.5	41.2	54.6	34.3	65.0	49.7	23.5	65.2	46.0	46.1	38.0	29.2	11.3
LLaMA-2-13B-Chat	41.0	31.1	43.2	25.6	35.7	73.0	45.5	25.0	69.9	49.4	27.1	39.0	32.2	22.6
ChatLaw-13B	50.9	35.6	46.7	44.8	38.0	51.5	61.5	55.9	67.5	42.4	14.7	45.3	24.6	23.7
LLaMA-2-7B-Chat	34.0	25.2	37.7	52.4	35.3	44.0	44.2	28.8	51.7	28.8	49.9	33.3	28.2	19.6
HanFei	45.6	25.1	48.6	42.7	33.5	40.6	32.5	25.1	76.3	55.1	23.4	40.1	34.3	19.5
MoSS-Moon-sft	34.1	27.5	33.7	35.5	32.1	36.0	38.2	29.7	62.1	26.3	4.5	34.0	26.4	20.0
Baichuan-7B-base	46.1	28.8	44.9	46.1	35.0	53.6	29.5	37.3	74.1	33.6	13.3	35.2	33.0	20.0
LLaMA-2-7B	36.3	24.7	27.3	33.0	29.6	38.0	56.9	26.1	38.8	26.9	25.2	30.1	7.4	35.9
MPT-7B	37.6	26.4	39.6	37.0	26.4	25.5	37.9	24.0	12.2	7.5	67.9	30.1	26.2	20.3
GoGPT2-13B	38.8	37.6	43.1	39.0	34.0	24.9	36.5	23.8	35.1	32.9	56.7	35.4	26.0	35.9
GoGPT2-7B	37.3	28.0	52.3	36.6	18.4	23.1	25.3	23.7	19.7	22.4	35.7	32.6	7.5	14.5
LaWGPT-7B-beta1.1	45.1	37.4	39.5	39.0	32.5	28.9	35.9	33.9	28.0	30.9	37.2	42.4	23.9	37.8
Alpaca-v1.0-7B	37.7	33.6	46.1	39.3	32.3	28.3	21.8	23.0	34.9	25.9	68.7	34.8	27.5	29.3
MPT-7B-Instruct	38.6	28.0	41.8	33.9	25.2	25.0	34.3	23.5	23.4	19.7	18.4	33.9	25.4	24.2
LaWGPT-7B-beta1.0	48.1	27.2	47.5	36.6	18.4	31.1	16.1	23.6	15.3	16.9	26.6	37.6	34.5	17.6
Vicuna-v1.3-7B	44.7	38.8	39.4	29.9	12.2	27.3	26.6	17.4	26.8	22.8	5.2	36.6	9.2	30.8
Lawyer-LLaMA	28.4	0.9	33.3	7.8	2.2	2.0	1.8	1.6	3.4	0.0	0.2	15.5	6.9	0.3
WisdomInterrogatory	9.4	17.9	6.1	1.3	5.8	3.5	0.0	0.2	4.2	3.0	0.1	1.2	8.2	9.4

Table 8: Zero-shot performance(%) of various models at Memorization, Understanding, and Logic Inference level. Mirco-F1, BERTScore, BARTScore are reported to provide a more comprehensive evaluation.

Model	Discrimination(F1)		Generation(BERTScore)				Generation(BARTSCore)				Ethic(F1)		
	4-1	4-2	5-1	5-2	5-3	5-4	5-1	5-2	5-3	5-4	6-1	6-2	6-3
ChatGLM	9.7	47.3	0.723	0.703	0.757	0.673	-3.98	-5.29	-7.16	-4.29	43.7	46.7	48.2
XVERSE-13B	12.3	47.2	0.657	0.678	0.596	0.686	-3.66	-5.74	-8.04	-4.74	63.0	64.6	73.5
Chinese-LLaMA-2-7B	14.3	42.9	0.698	0.654	0.756	0.634	-4.05	-9.06	-7.26	-4.94	43.3	47.1	52.0
Chinese-LLaMA-2-13B	4.3	44.5	0.698	0.677	0.766	0.627	-2.79	-5.15	-6.58	-3.75	38.8	41.1	41.9
Ziya-LLaMA-13B	38.6	45.1	0.702	0.672	0.779	0.684	-3.65	-6.03	-6.94	-4.60	37.2	44.9	43.2
LexiLaw	15.6	44.1	0.708	0.694	0.739	0.676	-4.19	-8.20	-7.01	-4.64	37.3	40.5	46.3
LLaMA-2-13B-Chat	35.0	52.3	0.669	0.624	0.711	0.654	-4.50	-8.84	-7.55	-4.60	37.5	45.9	48.6
ChatLaw-13B	27.3	60.2	0.673	0.603	0.834	0.657	-3.83	-6.29	-6.34	-4.08	42.9	51.9	51.1
LLaMA-2-7B-Chat	22.5	40.6	0.576	0.534	0.668	0.647	-4.87	-9.98	-7.77	-4.59	31.4	38.6	36.3
HanFei	2.6	37.4	0.683	0.697	0.742	0.679	-3.38	-5.28	-6.73	-4.37	41.2	43.8	43.4
MoSS-Moon-sft	16.5	41.5	0.695	0.717	0.785	0.672	-3.77	-4.83	-7.05	-4.27	30.5	38.6	38.2
Baichuan-7B-base	28.1	38.7	0.686	0.684	0.632	0.640	-2.90	-5.49	-6.93	-3.85	45.1	51.2	55.3
LLaMA-2-7B	23.5	35.3	0.703	0.600	0.599	0.616	-3.43	-10.35	-7.82	-5.14	33.8	42.4	34.0
MPT-7B	18.2	35.5	0.725	0.630	0.565	0.619	-3.07	-8.90	-7.98	-5.09	39.1	33.8	29.6
GoGPT2-13B	27.0	41.8	0.687	0.667	0.756	0.656	-3.97	-5.65	-7.44	-4.41	50.8	42.4	36.9
GoGPT2-7B	5.5	35.8	0.689	0.667	0.741	0.645	-3.60	-5.24	-7.57	-4.29	47.2	40.3	38.6
LaWGPT-7B-beta1.1	0.0	39.6	0.447	0.470	0.487	0.541	-5.97	-10.47	-7.64	-5.06	35.0	46.4	41.6
Alpaca-v1.0-7B	9.0	39.0	0.476	0.572	0.650	0.592	-6.53	-10.61	-7.84	-5.52	34.8	32.3	36.1
MPT-7B-Instruct	7.2	41.9	0.713	0.642	0.561	0.597	-3.20	-8.38	-7.83	-4.83	42.3	40.3	37.1
LaWGPT-7B-beta1.0	0.0	35.7	0.515	0.556	0.451	0.566	-5.13	-7.94	-6.72	-4.21	48.4	48.8	45.9
Vicuna-v1.3-7B	4.1	42.8	0.700	0.644	0.737	0.652	-4.15	-9.87	-7.22	-4.60	34.7	40.1	25.4
Lawyer-LLaMA	1.2	12.0	0.610	0.583	0.637	0.664	-4.50	-7.11	-7.18	-4.36	11.9	19.4	23.6
WisdomInterrogatory	25.0	10.4	0.686	0.699	0.768	0.659	-4.13	-4.92	-7.32	-3.70	8.4	15.9	20.0