Qzhou-Law: An Open Source Series of Chinese Legal Large Language Models

Yuchen Xie[1], Yixin Zhou[1], Huaitong Gao[1], Wensen Jiang[1], Jingxiang Fan[1], Chuhan Yan[1]
Zhiwei Fei[3], Yazhou Wang[1], Haibiao Chen[1], Yinfei Xu[1], Wei Zhou[2*]
Kingsoft AI, Kingsoft Corp. Ltd., Beijing, China[1] ,School of Law, Wuhan University, Wuhan, China[2], National Key Laboratory for Novel Software Technology, Nanjing University, China[3]

Abstract

We introduce a series of legal large language models (LLMs), Qzhou-Law 7B/14B/32B/72B. Our models achieve a new series of state-of-the-art (SOTA) performances on LawBench, LexEval and NULPQE, which is construted to evaluate the timely changes to laws, regulations and relevant knowledge. The core innovations are that (1) we construct a large-scale legal instruction-tuning dataset and (2) we explore a new training method to train a legal-specific LLM in three phases better.

Objectives

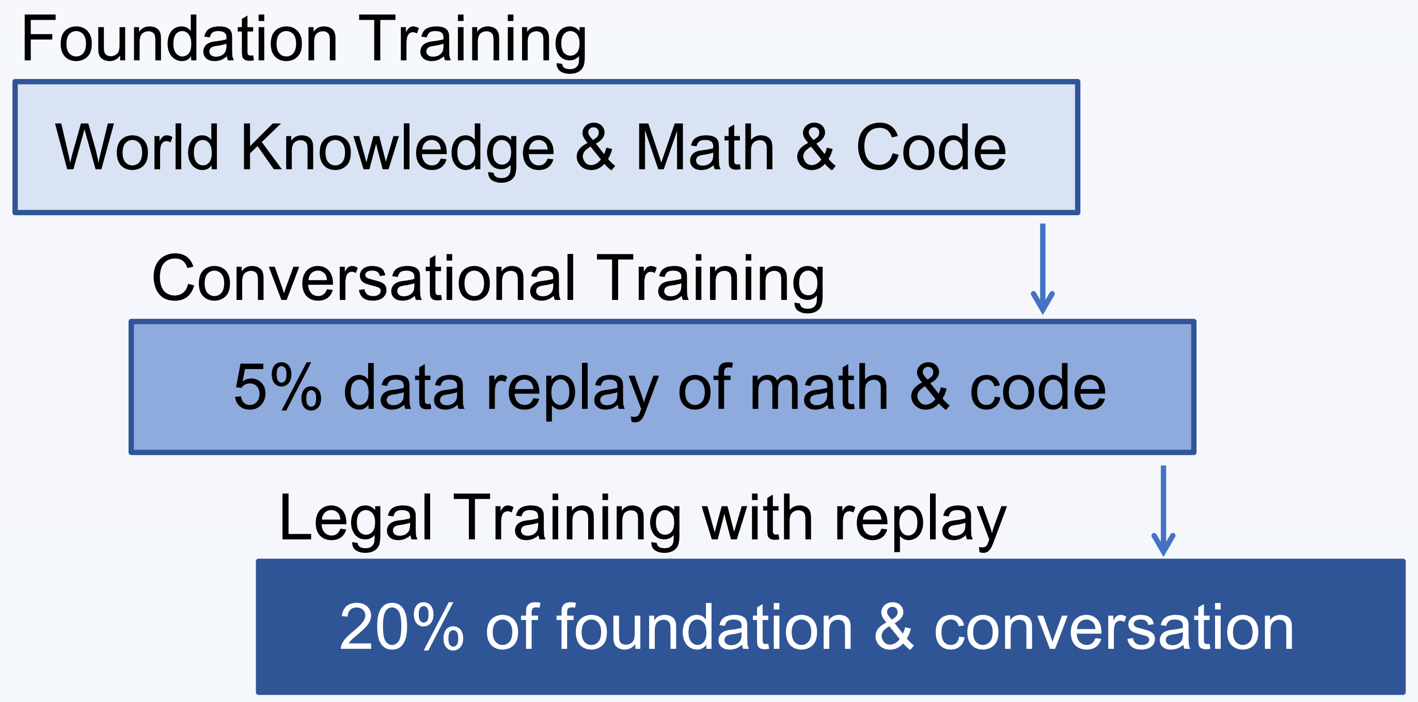
- Propose a novel three-stage fine-tuning approach to train a legal LLM with a proper data replay strategy.
- Construct a large-scale legal dataset including legal scenarios, legal knowledge, counseling, laws, and regulations.
- Establish the NULPQE benchmark to evaluate timely changes in laws.

Methods

- Stage 1: Train base models on Foundation datasets of Infinity-Instruct
- Stage 2: Enhance chat abilities with a 5% data replay strategy (math & code) to retain foundational skills.
- Stage 3: Fine-tune on legal datasets mixing 20% general data to maintain general capabilities.

Methods

Three-Stage Legal Instruct Training



Foundation Training

World Knowledge & Math & Code

Conversational Training

5% data replay of math & code

Legal Training with replay

20% of foundation & conversation

High-Quality Data Construction

To train Qzhou-Law, we construct a comprehensive dataset containing over 853K training samples.

Dataset	Source	Size
Legal Knowledge	JEC-QA	25K
	Website	55K
Laws and Regulations	Website	100K
Legal Counseling	Hanfei	42K
	DISC-Law	23K
	LawGPT	35K
	Lawyer-llama	1K
	Website	255K
Legal Scenarios	Public Datasets	317K

Results

Our models achieve SOTA performances with scaling trend on LexEval.

Results

Level	GPT-5.1	DeepSeek-V3	InternLM-Law	Qzhou-Law-72B
Memorization	43.14	50.98	49.07	71.3
Understanding	83.09	85.53	77.81	87.92
Logic Inference	66.13	70.03	61.79	83.08
Discrimination	30.74	30.07	32.67	36.44
Generation	21.56	24.76	26.46	45.42
Ethic	56.3	58.1	59.67	70.73
AVG	54.71	58.53	54.66	70.38

Level	Qzhou-Law-7B	Qzhou-Law-14B	Qzhou-Law-32B	Qzhou-Law-72B
Memorization	55.44	64.94	66.68	71.30
Understanding	77.88	84.39	89.36	87.92
Logic Inference	72.26	79.22	80.92	83.08
Discrimination	22.97	27.87	32.77	36.44
Generation	41.28	43.76	38.00	45.42
Ethic	61.83	64.97	68.70	70.73
AVG	60.25	65.99	67.65	70.38

Our models demonstrate strong scaling trends and putperform on LawBench.

Level	GPT-5.1	DeepSeek-V3	InternLM-Law	Qzhou-Law-72B
Memorization	41.19	65.99	64.34	76.09
Understanding	52.59	48.52	71.70	76.15
Application	62.10	64.05	63.52	75.27
AVG	55.27	56.48	67.69	75.79

Results

Level	Qzhou-Law-7B	Qzhou-Law-14B	Qzhou-Law-32B	Qzhou-Law-72B
Memorization	62.27	67.15	72.30	76.08
Understanding	71.35	73.82	73.39	76.15
Application	71.65	71.72	74.50	75.27
AVG	70.56	72.31	73.75	75.79

To evaluate the timely changes in laws, regulations, and relevant knowledge, we developed the NULPQE benchmark and our models outperform competitors.

Year	GPT-5.1	DeepSeek-V3	Law-LLM-7B	Qzhou-Law-7B	Qzhou-Law-72B
2024	111	113	121	154	193
2023	147	154	165	172	227
2022	132	150	108	143	202
2021	107	127	93	167	180
2020	121	141	132	179	242
2019	144	146	141	189	219
2018	145	133	130	204	241
AVG	130	138	127	173	215

Conclusion

- We trained a series of models achieving a new SOTA performances on LawBench, LexEval, and NULPQE.
- We demonstrated strong scaling trends of performance on most tasks in LawBench and LexEval.