AppealCase: A Dataset and Benchmark for Civil Case Appeal Scenarios

Yuting Huang[1], Meitong Guo[1], Yiquan Wu[1], Ang Li[1], Mengze Li[1]
Xiaozhong Liu[2], Keting Yin[1], Changlong Sun[1], Kun Kuang[1*]

[1]Zhejiang Univeristy, Hangzhou, China
[2]Worcester Polytechnic Institute, Worcester, USA

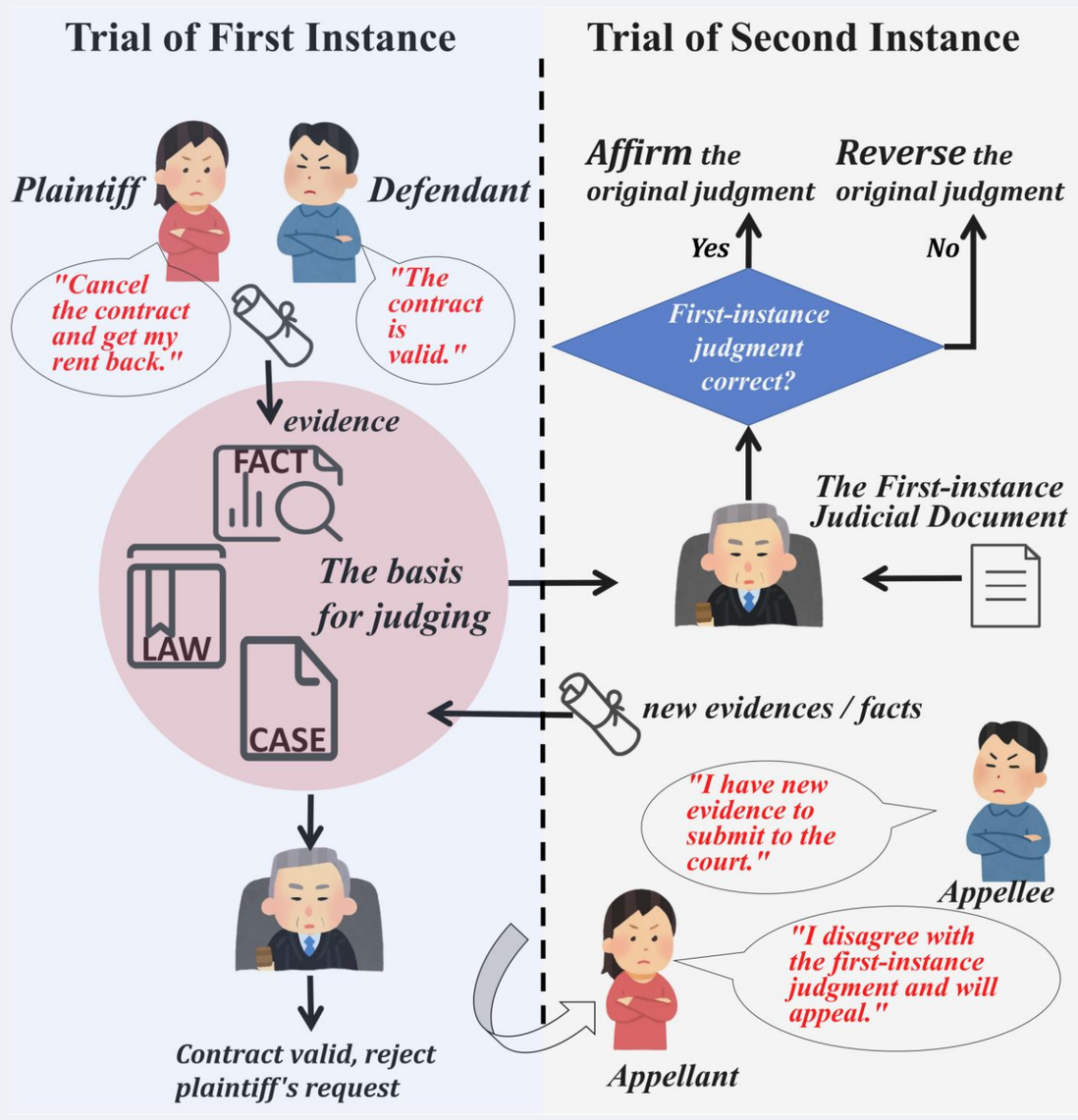
https://github.com/ythuang02/AppealCase

https://huggingface.co/datasets/ythuang02/AppealCase

Introduction

The **AppealCase** dataset is the first large-scale resource specifically designed to support LegalAI research in **appellate judgment scenarios**.

While prior work in LegalAI has focused heavily on one-shot trials, the appellate procedure, which is critical to ensuring fairness and correcting judicial errors, remains largely underexplored.



AppealCase Dataset

- **Cases**: **10,000 matched pairs** of first-instance and second-instance judgments
- **Coverage**: **91 civil causes of action**
- **Format:** JSON structured format
- **License:** CC BY-NC 4.0

Annotation Scheme

- **Judgment Reversal:** A binary label indicating whether the second-instance court overturned the first-instance decision.
- **Reasons for Reversal:** This includes *errors in factual determination* and *errors in the application of law*.
- **Claims:** A list of individual claims raised in the first-instance proceedings.
- **Legal Provisions:** A list of legal provisions explicitly cited in the second-instance judgmen
- **New Information:** A binary label indicating whether newevidence were introduced during the appeal.

Tasks

- **Judgment Reversal Prediction (first-instance):** Given the first-instance document and the second-instance claim to predict the reasons for reversal.
- **Judgment Reversal Prediction (second-instance):** Given the first-instance document and the second-instance claim and fact description, which contains new information introduced in the second instance, the task is to predict the reasons for reversal.
- **Provision Relevance Prediction**
- **Legal Judgment Prediction**
- **Judgment Reversal Prediction**

Results on Judgment Reversal Prediction

Category	Model	First-instance Perspective			Second-instance Perspective		
		Precision	Recall	F1	Precision	Recall	F1
Non-Reasoning	DeepSeek-V3	44.94	41.87	42.53	55.56	54.97	54.49
	Qwen2.5-72B	47.62	41.23	40.49	55.84	59.45	**57.40**
	LLaMA3.3-70B	40.75	45.22	34.85	50.42	56.94	48.08
	GPT-4.1	**51.93**	36.53	32.58	**60.28**	47.30	44.80
	GLM-4-Air	38.64	42.09	33.24	42.08	41.10	38.72
	Doubao-1-5-pro	42.57	47.42	**44.57**	55.28	59.27	54.73
	Baichuan2-7B	26.00	33.37	26.60	34.99	34.87	25.90
	Qwen2.5-7B	38.28	34.31	30.42	46.02	41.56	40.18
	Llama3.1-8B	34.38	34.34	18.43	41.45	36.06	28.05
Reasoning	DeepSeek-R1	44.17	43.54	43.06	54.03	55.56	54.77
	R1-Distill-Qwen-32B	40.01	45.61	40.58	49.28	57.36	52.07
	QwQ-32B	42.31	**51.41**	40.30	52.38	54.79	49.95
	Qwen3-32B	40.06	49.19	39.30	49.91	59.19	50.64
	GLM-Z1-Air	39.34	43.34	39.46	49.98	54.34	48.90
	GPT-o4-mini	43.67	40.49	40.36	54.31	49.16	47.85
	Grok-3-mini	37.23	49.37	37.41	48.97	**62.64**	53.69
	R1-Distill-Qwen-7B	35.82	44.84	37.71	37.87	52.98	41.68
	Qwen3-8B	39.79	51.34	36.03	49.21	55.44	47.35
Domain	DISC-LawLLM	32.51	33.85	30.05	35.11	34.11	23.09
	Wisdom Interrogatory	32.92	34.20	30.47	33.74	33.93	22.52

- **All models perform poorly on the judgment reversal prediction task**, highlighting its difficulty. Under the first-instance perspective, no model achieves an F1 score above 50%; under the second-instance perspective, more than half of the models remain below 50%.
- **Existing domain-specific models are constrained by limited context windows** and struggle to process long, structured judicial documents.
- Models perform better in the second-instance perspective, likely due to the inclusion of summarized information from the first-instance trial, which aids reasoning.