CoreEval: Automatically Building Contamination-Resilient Datasets with Real-World Knowledge toward Reliable LLM Evaluation

Jingqian Zhao[1*], Bingbing Wang[1*], Geng Tu[1], Yice Zhang[1], Qianlong Wang[1], Bin Liang[4†], Jing Li[5], Ruifeng Xu[1,2,3†]
[1] Harbin Institute of Technology, Shenzhen, China [2] Peng Cheng Laboratory, Shenzhen, China
[3] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[4] The Chinese University of Hong Kong, Hong Kong, China
[5] The Hong Kong Polytechnic University, Hong Kong, China

@HITSZ

ACL 2025 VIENNA
JULY 27 - AUGUST 1

Introduction

Task Definition: The task is to address data contamination in Large Language Model (LLM) evaluation. The static nature of public benchmarks leads to test data being inadvertently included in training sets, which artificially inflates performance and compromises the reliability and fairness of evaluations.

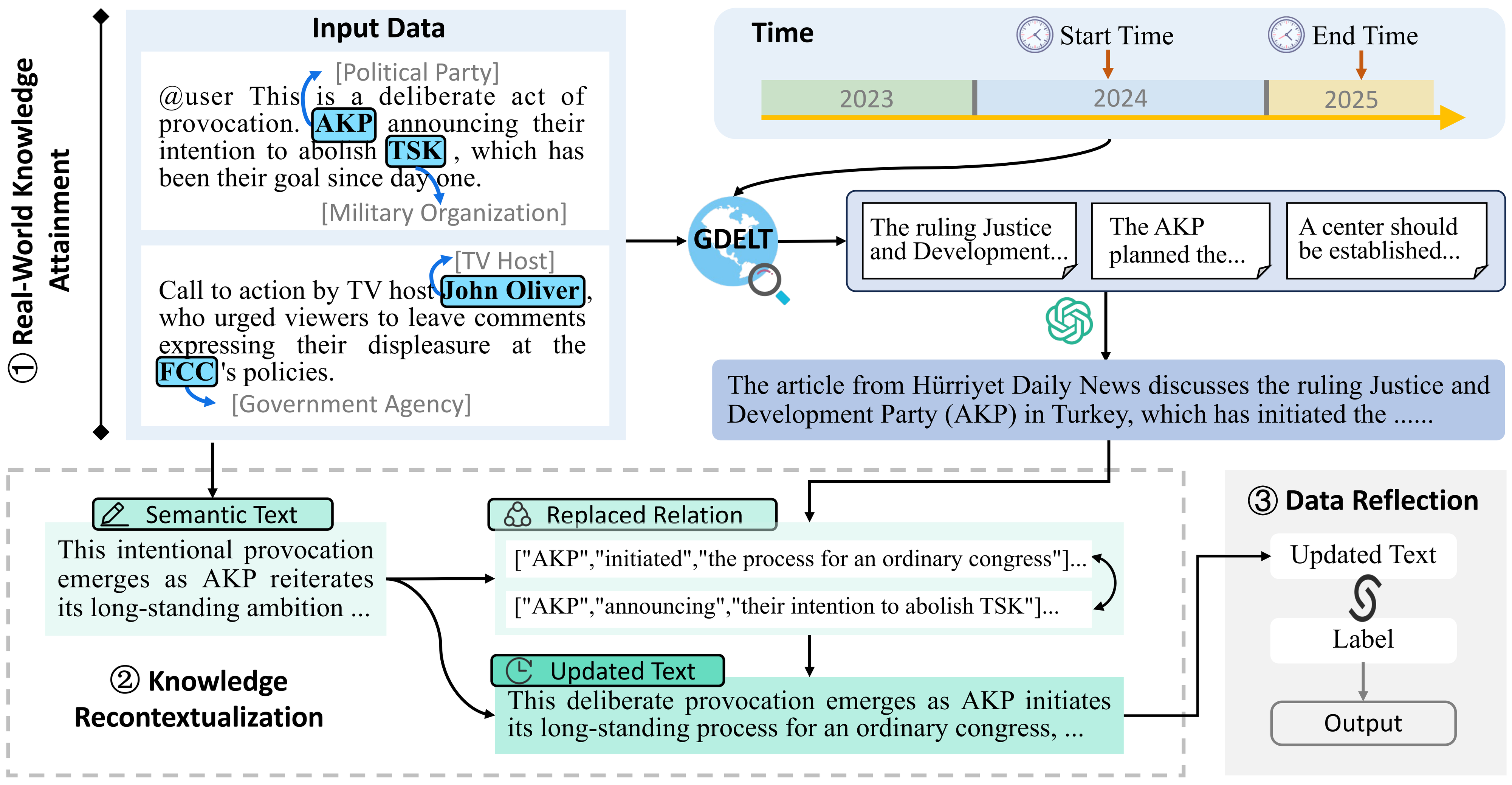
Motivation: Current automated methods to create new datasets are insufficient. Data rewriting risks producing inconsistent labels and re-introducing contamination from the model's own biases , while data generation often fails to preserve the original dataset's semantic complexity, leading to information loss.

Contribution

- We propose CoreEval, an automatic contamination-resilient evaluation strategy that integrates real-world knowledge to update datasets.
- We design a structured workflow inspired by cognitive learning theory to ensure reliable and timely LLM evaluation.
- Extensive experiments across multiple tasks and a series of LLMs demonstrate the effectiveness of CoreEval in mitigating data contamination.

CoreEval Framework

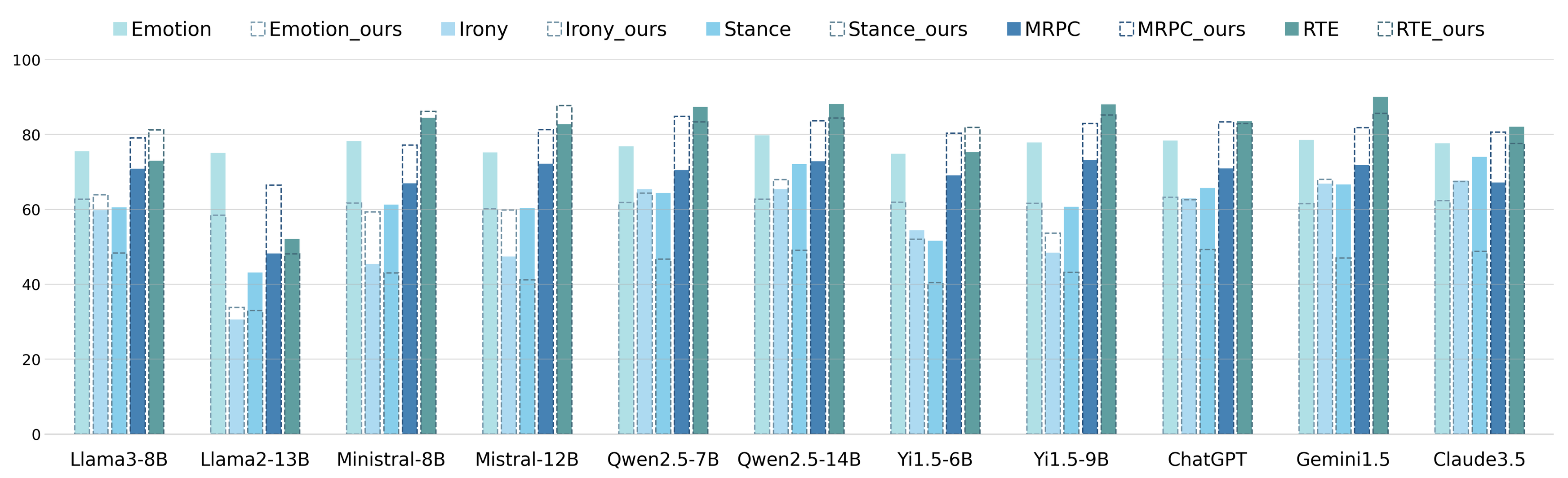
1. **Real-World Knowledge Attainment:**
   - **Entity Extraction:** $E_i \leftarrow \mathcal{M}(d_i)$
     - $\mathcal{M}$: Large Language Model
     - d_i: source data
     - E_i: set of extracted entities
   - **Knowledge Retrieval:** $\mathcal{K}_i \leftarrow \mathcal{G}(E_i, t_{\text{start}}, t_{\text{end}})$
     - $\mathcal{G}$: GDELT Database
     - $\mathcal{K}_i$: retrieved knowledge
   - **Knowledge Summary:** $\hat{\mathcal{K}}_i \leftarrow \mathcal{M}(\mathcal{K}_i)$
     - $\hat{\mathcal{K}}_i$: summarized knowledge
2. **Knowledge Recontextualization**
   - **Triple Update:**
     - Extract original triples T_i from d_i.
     - Generate updated triples: $\hat{T}_i \leftarrow \mathcal{M}(T_i, \hat{\mathcal{K}}_i)$.
   - **Text Synthesis:**
     - Create content-updated text: $d_i^u \leftarrow f(d_i, \hat{T}_i)$.
     - Create style-preserving text: $d_i^s \leftarrow \mathcal{M}(d_i, T_i)$.
   - **Final Integration:** $\hat{d}_i \leftarrow \mathcal{M}(d_i, d_i^u, \hat{T}_i, d_i^s)$.



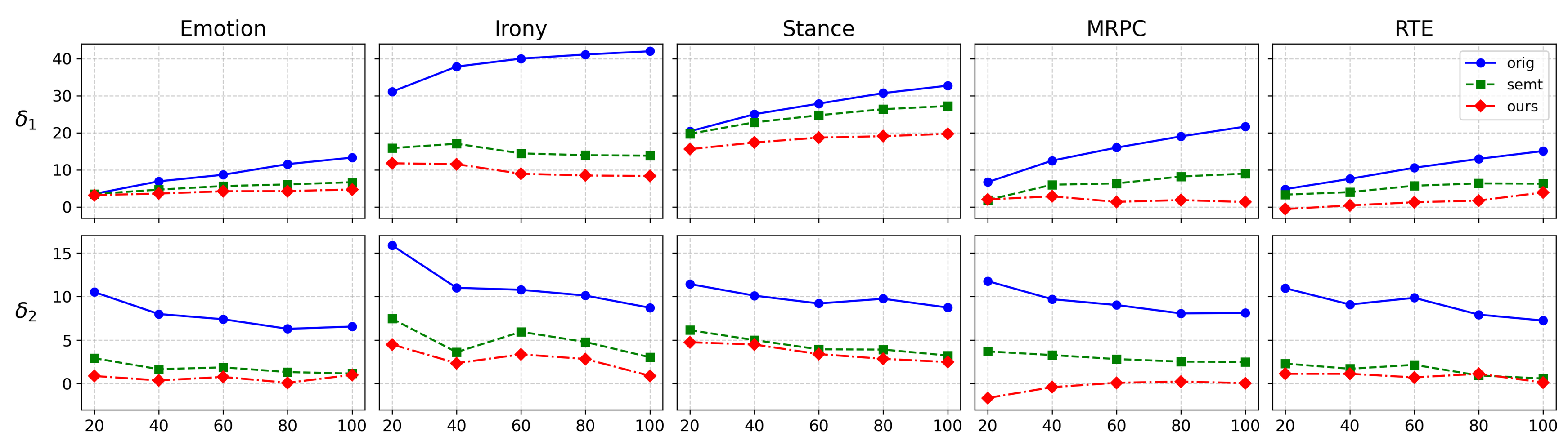
3. **Data Reflection**

   An agent evaluates $\hat{d}_i$ via prompting. Iteratively re-generate $\hat{d}_i$ if checks fail.

Experimental Results & Conclusion



Performace Test



Impact of Contamination Proportion

Main Experimental Results:

		Emotion		Irony		Stance		MRPC		RTE		AVG	
		$\delta_1\downarrow$	$\delta_2\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$
Llama3-8B	*orig*	9.37	4.47	30.09	7.07	23.41	6.80	10.98	7.05	20.88	8.79	18.95	6.84
	semt	4.86	1.34	9.66	3.05	20.00	3.14	6.37	2.78	12.20	**0.12**	10.62	2.09
	ours	**3.27**	**1.33**	**2.00**	**1.89**	**11.66**	**2.57**	**0.75**	**0.53**	**4.21**	0.13	**4.38**	**1.29**
Llama2-13B	*orig*	11.83	4.55	52.46	7.97	38.26	6.98	26.98	6.83	26.24	9.02	31.16	7.07
	semt	7.69	1.60	23.15	2.42	32.70	3.26	18.44	2.50	22.63	2.15	20.92	2.38
	ours	**7.41**	**0.57**	**18.23**	**1.12**	**24.50**	**2.56**	**10.09**	**-0.31**	**21.12**	**1.10**	**16.27**	**1.01**
Ministral-8B	*orig*	12.65	6.85	39.66	7.58	30.80	8.51	25.64	8.91	17.08	6.64	25.17	7.70
	semt	6.77	1.58	10.53	1.98	28.47	3.38	11.94	2.84	6.50	-0.72	12.84	1.81
	ours	**4.41**	**0.15**	**2.54**	**0.58**	**20.97**	**2.36**	**3.86**	**0.32**	**4.03**	**-1.46**	**7.16**	**0.39**
Mistral-12B	*orig*	17.41	7.59	40.43	10.59	34.69	9.35	26.51	9.30	13.43	7.49	26.50	8.86
	semt	10.83	**1.44**	8.46	4.27	30.49	3.69	10.54	2.28	2.74	0.11	12.61	2.36
	ours	**7.64**	1.54	**2.92**	**3.40**	**23.35**	**3.19**	**0.61**	**0.43**	**1.45**	**-0.62**	**7.19**	**1.59**
Yi1.5-6B	*orig*	11.42	4.65	39.78	8.45	31.75	8.64	14.96	7.71	19.64	8.80	23.51	7.69
	semt	4.76	**0.60**	20.47	2.62	24.70	2.46	6.92	1.39	11.16	**0.36**	13.60	1.49
	ours	**3.50**	0.84	**16.35**	**0.75**	**18.79**	**2.45**	**-1.00**	**0.21**	**6.76**	1.69	**8.88**	**1.19**
Yi1.5-9B	*orig*	15.03	9.04	44.34	14.13	33.67	11.60	23.87	10.41	9.21	8.08	25.22	10.65
	semt	6.17	1.94	12.86	2.31	25.50	3.79	6.48	1.89	2.53	1.71	10.71	2.33
	ours	**4.50**	**0.51**	**7.59**	**0.55**	**19.66**	**2.00**	**-3.50**	**0.27**	**0.45**	**-0.37**	**5.74**	**0.59**
Qwen2.5-7B	*orig*	6.65	3.44	19.77	4.86	18.51	5.24	8.06	4.16	7.08	6.03	12.01	4.74
	semt	4.93	1.06	10.37	2.77	18.06	2.87	3.21	2.32	**1.74**	**0.72**	7.66	1.95
	ours	**4.72**	**0.61**	**6.82**	**2.31**	**15.25**	**2.32**	**0.04**	**-0.39**	2.31	1.08	**5.83**	**1.19**
Qwen2.5-14B	*orig*	11.53	5.71	27.75	9.78	20.83	8.10	19.95	6.94	7.21	5.43	17.45	7.19
	semt	5.79	1.46	1.46	2.76	17.03	2.87	5.49	1.42	**0.52**	0.12	6.06	1.72
	ours	**4.57**	**0.99**	**-3.57**	**0.93**	**13.98**	**1.20**	**-4.73**	**0.37**	4.38	**0.00**	**2.93**	**0.70**

Dataset Statistics & Quality Evaluation:

Dataset	Train	Test	Label Space
Emotion	3,257	1,421	joy, optimism, sadness, anger
Irony	2,862	784	irony, not irony
Stance	2,620	1,249	favor, against, neutral
MRPC	4,076	1,587	equivalent, not equivalent
RTE	2,490	277	entailment, not entailment

Dataset	Fluency	Coherence	Factuality	Accuracy	κ
Emotion	2.99	2.55	0.98	0.94	0.73
Irony	2.97	2.74	0.99	0.97	0.78
Stance	2.99	2.56	0.98	0.96	0.73
MRPC	2.98	2.92	0.98	0.96	0.86
RTE	2.99	2.86	0.96	0.96	0.80

Performance on Updated Data

- LLM performance drops significantly on the updated datasets, especially for subjective tasks like stance detection, suggesting original benchmarks are contaminated.
- Proprietary models show a larger performance drop than open-source models (5.42% vs 3.62%), implying more severe contamination.

Contamination Resistance

- In simulated contamination scenarios, CoreEval's updated dataset shows significantly stronger resistance to performance overestimation than original and rewritten datasets.
- The framework effectively mitigates the impact of contamination across different model sizes and varying contamination proportions (20%-100%).

Contact Us

Authors

Jingqian Zhao: `zhaojingqian@stu.hit.edu.cn`
Bingbing Wang: `bingbing.wang@stu.hit.edu.cn`

About

Jingqian Zhao: https://zhaojingqian.top/about

For collaboration or opportunities, feel free to contact me.

WeChat

CoreEval