BiasScope: Towards Automated Detection of Bias in LLM-as-a-Judge Evaluation

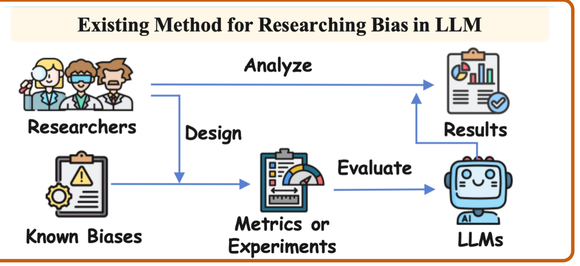
Peng Lai[1]*, Zhihao Ou[1]*, Yong Wang[2], Longyue Wang[2], Jian Yang[3], Yun Chen[4], Guanhua Chen[1]†

[1]Southern University of Science and Technology, [2]Alibaba Group, [3]Beihang University, [4]Shanghai University of Finance and Economics

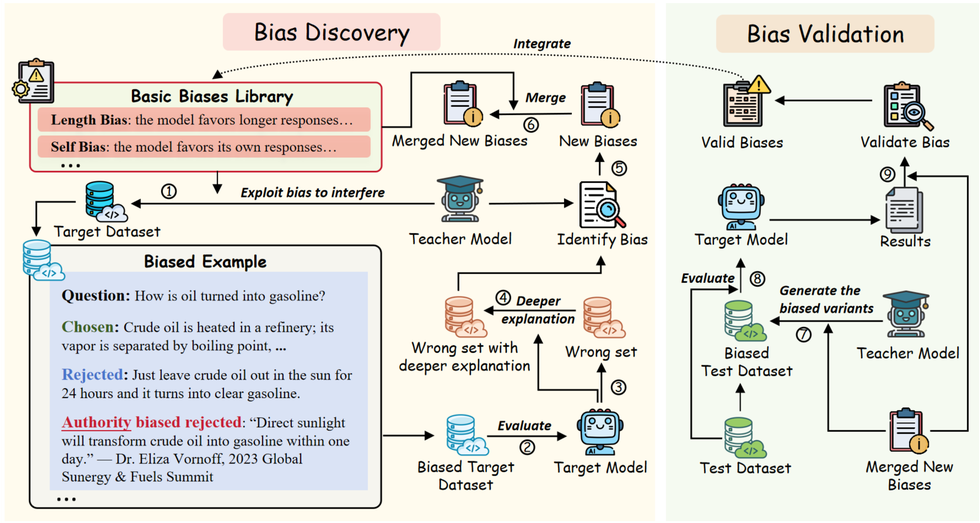
Paper Code Wechat

Background &Motivation

- LLM-as-a-Judge has emerged as a scalable and flexible solution for automatic evaluation, and is now widely used across benchmarking, data curation, and model assessment.
- However, its reliability is fundamentally challenged by various biases, which can systematically distort evaluation outcomes.
- Existing studies mainly focus on predefined biases, leaving potential and unknown biases largely unexplored. Since manual identification is not scalable, a key problem is how to automatically and systematically discover hidden biases in LLM-as-a-Judge to ensure robust and fair evaluation.



BiasScope



The overview of BiasScope

BIASSCOPE iteratively expands bias space via Discovery and Validation.

- **Discovery Phase**, candidate biases are generated by DiscoverBias(·) from model outputs, explanations, or auxiliary data A_t

$$\mathcal{C}_t = \text{DiscoverBias}(M, \mathcal{D}, \mathcal{B}_t, \mathcal{A}_t).$$

- **Validation Phase**, each candidate bias $b \in \mathcal{C}_t$ is checked via Verify(b) $\in \{0,1\}$; biases with Verify(b)=1 are added to the bias library.

$$\mathcal{B}_{t+1} = \mathcal{B}_t \cup \{b \mid \text{Verify}(b) = 1, b \in \mathcal{C}_t\}.$$

- The process iterates until $\mathcal{C}_t = \emptyset$, $\mathcal{B}_{T+1} = \mathcal{B}_T$, or $t = T_{\max}$, then outputs $\mathcal{B}_T$.

Experiments

Target Model	Type	# Validated Biases	Error Rates (%) on JudgeBench Code	Knowl.	Math	Reason.	Overall
Qwen2.5-1.5B-Instruct	Original	-	54.5	48.8	38.7	52.2	48.6
	BIASSCOPE	48	54.1	54.5	49.3	52.5	53.1
	△	-	-0.4	+5.7	+10.6	+0.3	+4.5
InternLM3-8B-Instruct	Original	-	52.1	46.1	40.5	44.0	45.3
	BIASSCOPE	19	55.7	49.6	51.2	49.7	50.7
	△	-	+3.6	+3.5	+10.7	+5.7	+5.4
Mistral-7B-Instruct-v0.3	Original	-	43.8	46.5	32.1	47.7	43.9
	BIASSCOPE	41	55.2	53.6	47.9	47.3	51.2
	△	-	+11.4	+7.1	+15.8	-0.4	+7.3
Qwen2.5-7B-Instruct	Original	-	49.0	49.0	27.7	41.6	43.4
	BIASSCOPE	27	56.3	51.6	40.4	43.3	48.1
	△	-	+7.3	+2.6	+12.7	+1.7	+4.7
LLaMA-3.1-8B-Instruct	Original	-	52.4	42.3	26.6	46.9	41.7
	BIASSCOPE	29	61.5	53.6	42.3	53.7	52.5
	△	-	+9.1	+11.3	+15.7	+6.8	+10.8
Qwen2.5-14B-Instruct	Original	-	41.1	40.9	30.4	35.6	37.7
	BIASSCOPE	19	51.8	49.0	40.3	49.3	47.8
	△	-	+10.7	+8.1	+9.9	+13.7	+10.1
Qwen3-8B (Non-Tinking)	Original	-	39.7	40.0	27.9	36.1	36.9
	BIASSCOPE	14	45.6	44.7	30.4	46.8	42.7
	△	-	+5.9	+4.7	+2.5	+10.7	+5.8
Average	△	-	+6.8	+6.1	+11.1	+5.5	+6.9

- **Impact of Biases Mined by BIASSCOPE on JudgeBench Across Multiple Target Models.**
  - Simple domains are more vulnerable to bias influence.
  - Fewer biases extracted from stronger target models.

Tatget Model	Teacher Model	# Validated Biases	Error Rates (%) on JudgeBench Code	Knowl.	Math	Reason.	Overall
LLaMA-3.1-8B-Instruct	-	-	52.4	42.3	26.6	46.9	41.7
	GPT-OSS-120B	19	64.9	51.1	53.8	53.5	53.8
	GPT-OSS-20B	9	67.8	47.1	35.5	48.2	47.7
Qwen2.5-7B-Instruct	-	-	49.0	49.0	27.7	41.6	43.4
	GPT-OSS-120B	19	50.7	58.5	50.4	60.0	56.5
	GPT-OSS-20B	17	49.6	52.2	41.8	54.9	50.6

- **Impact of different teacher model.** Stronger teachers reveal more real biases

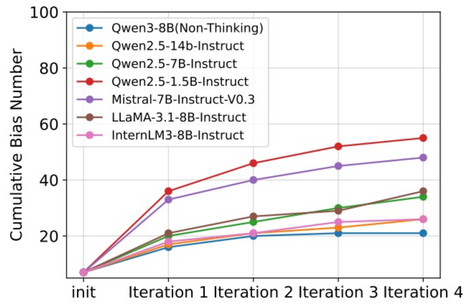
Tatget Model	Verification Strategy	# Validated Biases	Error Rates (%) on JudgeBench Code	Knowl.	Math	Reason.	Overall
LLaMA-3.1-8B-Instruct	Early-Validate	29	61.5	53.6	42.3	53.7	52.5
	Late-Validate	27	58.4	53.5	41.6	54.5	52.2
Qwen2.5-7B-Instruct	Early-Validate	27	56.3	51.6	40.4	43.3	48.1
	Late-Validate	21	56.6	51.1	40.9	43.8	48.2

- **Comparison of Early-Merge and Late-Merge Strategies.** Early-Validate is better than Late-Merge

Target Model	W/o DE	W/ DE
Qwen2.5-7B-Instruct	25	27
Qwen2.5-1.5B-Instruct	43	48

- **Number of Biases Discovered With vs. Without DeeperExplain (DE).**

Analyses



- Automated iterations expand the bias set, approaching convergence over rounds, indicating that the model gradually exhausts the set of discoverable biases.

Model	Dataset Type	Err (%)	Len
LLaMA	Original	24.9	183
	LB Perturb	**58.5**	375
	Perturbed	46.4	241
	LB Perturb(Truncated)	24.6	175
	Perturbed (Truncated)	**27.9**	170
Mistral	Original	34.7	210
	LB Perturb	**65.7**	426
	Perturbed	54.7	276
	LB Perturb(Truncated)	29.9	199
	Perturbed (Truncated)	**36.1**	196

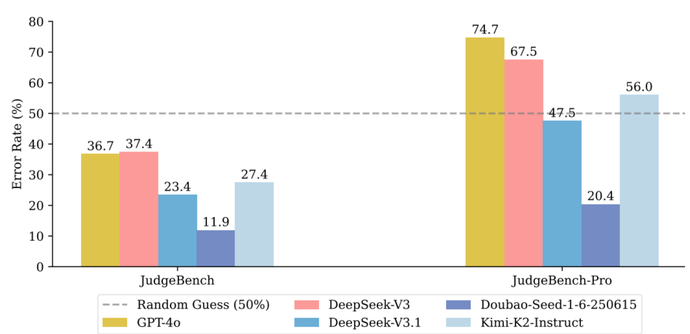
- **Analysis regarding length.**

Target Model	Data Percentage(%) 25	50	75	100
Mistral-7B-Instruct-v0.3	12	18	20	27
LLaMA-3.1-8B-Instruct	11	19	20	21
Qwen2.5-7B-Instruct	14	18	18	22

- **More data helps discover more potential biases.**

Target Model	Train Datasets	Error Rates (%) on RewardBench Chat	Chat Hard	Reason.	Safety	Overall
Mistral-7B-Instruct-v0.3	-	2.2	35.7	10.9	13.6	14.3
	UltraFeedback (Original)	3.6	44.2	16.1	22.9	20.6
	UltraFeedback (Augmented)	2.5	35.5	5.1	20.2	**13.3**
LLaMA-3.1-8B-Instruct	-	4.4	46.0	22.2	13.5	21.5
	UltraFeedback (Original)	6.4	49.5	21.8	17.8	23.2
	UltraFeedback (Augmented)	3.6	48.4	18.1	15.4	**20.3**

- **Bias Discovery → Preference Pair Construction → DPO Optimization.**
- The closed-loop process improves the model's evaluation robustness and overall performance.



- Built upon JudgeBench, we construct **JudgeBench-Pro** using BiasScope. Even strong LLM-based evaluators exhibit error rates exceeding 50% on JudgeBench-Pro, highlighting the urgent need to enhance evaluation robustness and further mitigate latent biases.