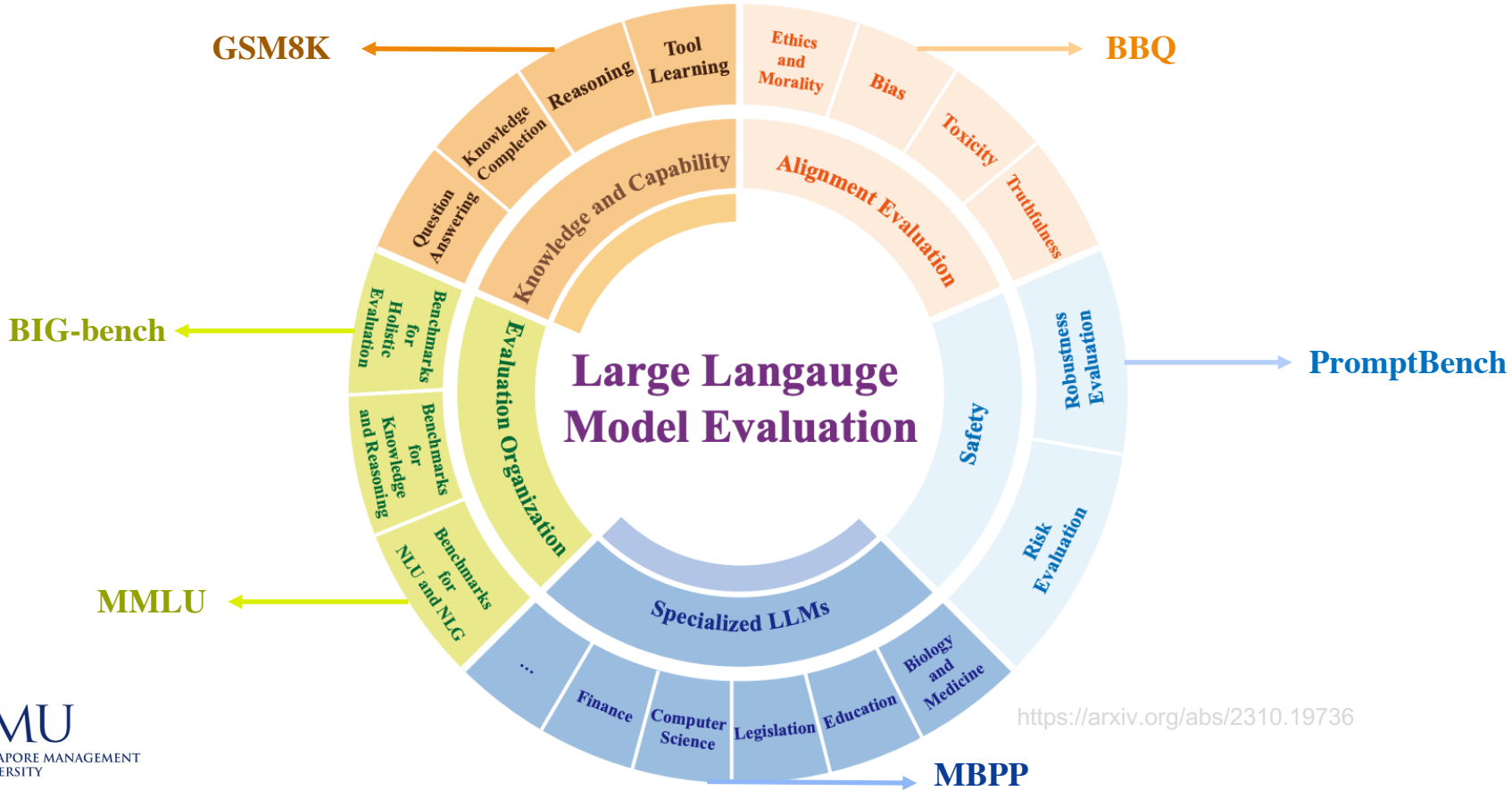
SMU
SINGAPORE MANAGEMENT UNIVERSITY
School of
Computing and
Information Systems

Automating Dataset Updates Towards Reliable and Timely Evaluation of Large Language Models

Jiahao Ying, Yixin Cao, Yushi Bai, Qianru Sun, Bo Wang, Wei Tang,
Zhaojun Ding, Yizhe Yang, Xuanjing Huang, Shuicheng Yan

Currently work for Evaluation



https://arxiv.org/abs/2310.19736

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Motivation:

- Data Leakage: Ensuring the Reliability of evaluation poses a significant challenge.

1. *"Skywork: A More Open Bilingual Foundation Model"*
2. *"An Open Source Data Contamination Report for Llama Series Models"*
3. *"Don't Make Your LLM an Evaluation Benchmark Cheater"*

Generate new clean test samples and update the exiting datasets

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Motivation:

- Effectively distinguish: LLMs are gradually mastering more challenging datasets

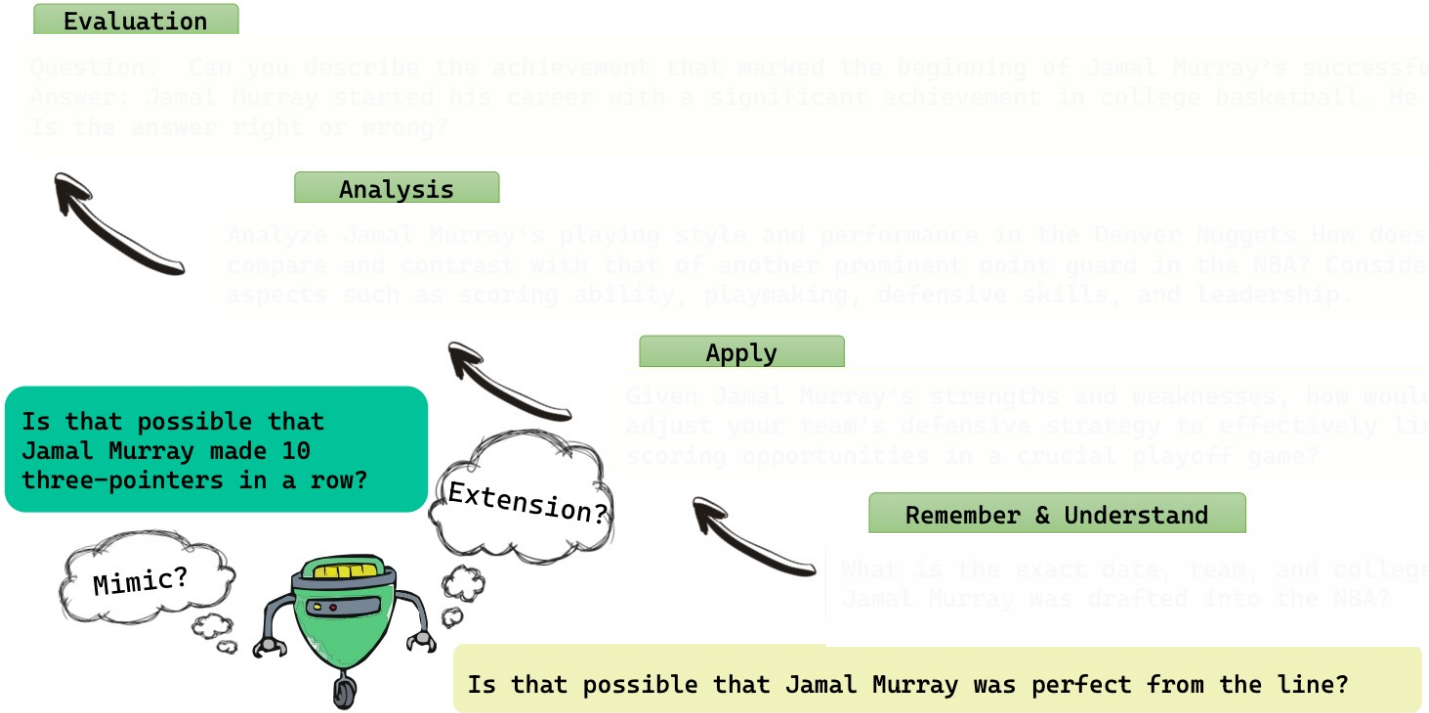
Clearly, such a way of manually constructing constantly updating dataset is costly

Demanding experts to curate more difficult datasets

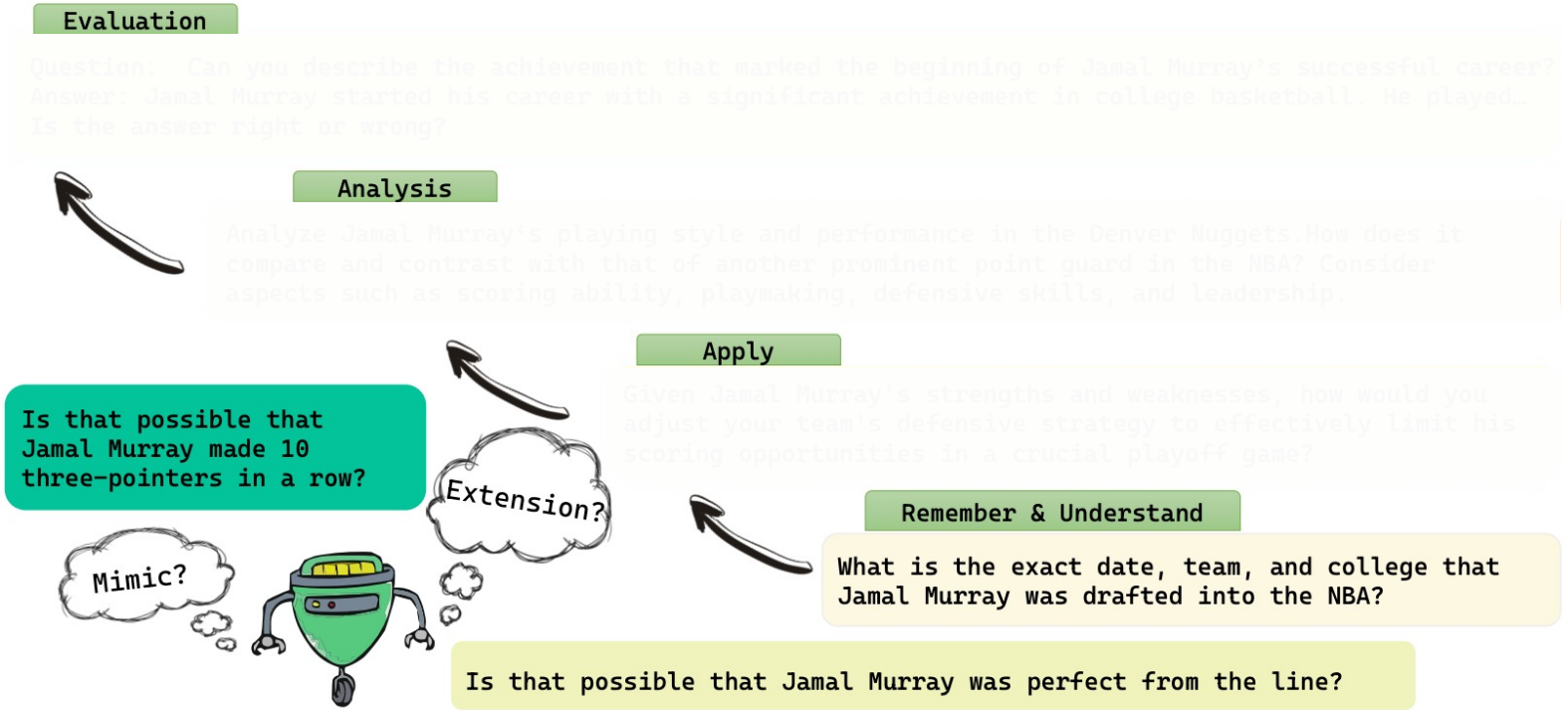
Motivation

In this work, we aim to **Automate updating benchmarks for LLMs evaluation,** minimizing human efforts, achieving timely and reliably evaluation.

Auto-updating Framework



Auto-updating Framework



Motivation

In this work, we aim to **Automate updating benchmarks for LLMs evaluation,** minimizing human efforts, achieving timely and reliably evaluation.

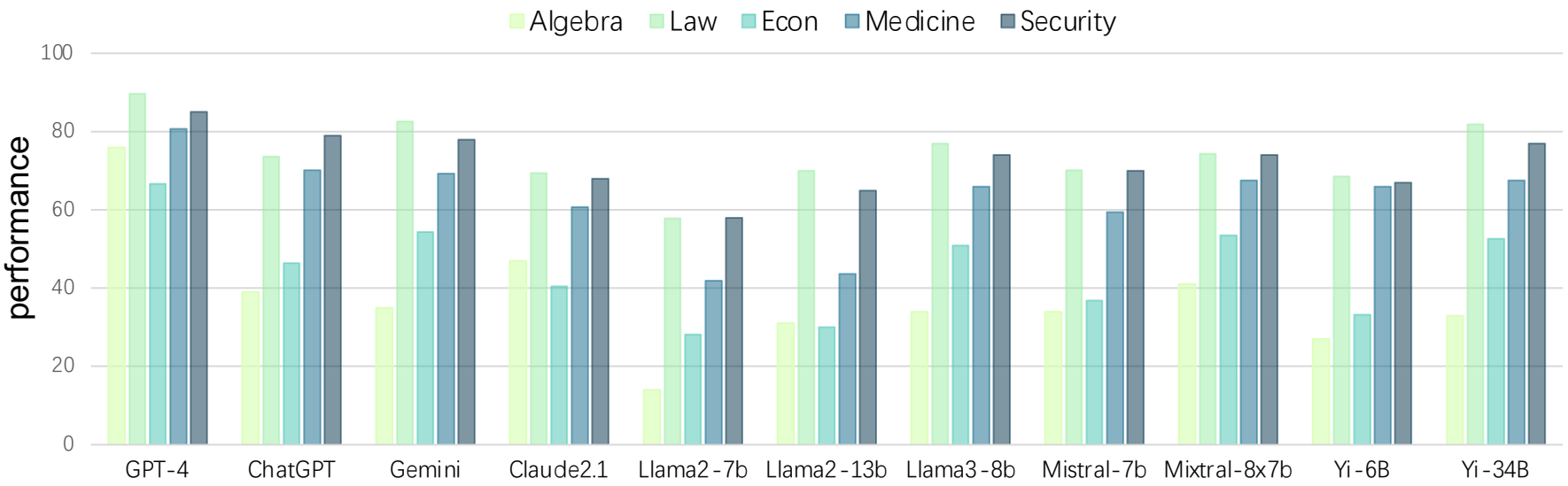
However, this is non-trivial due to the following research questions:

- **Stability:** Will updated benchmarks produce stable results?

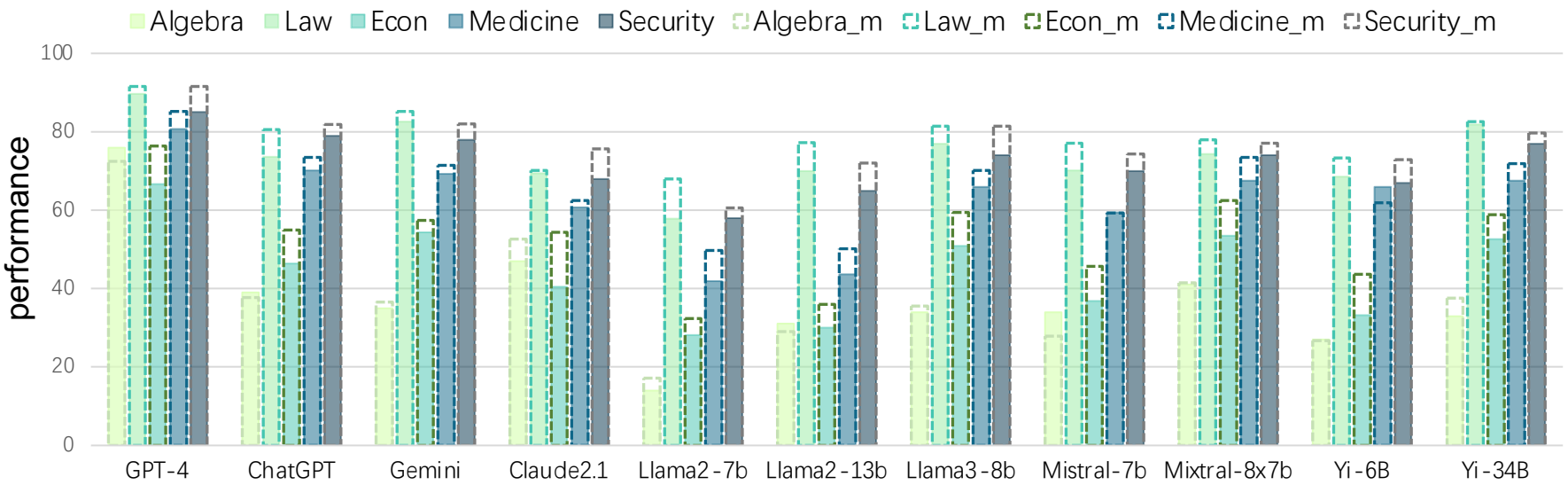
- **Reliability:** How can the update strategy mitigate benchmark leakage issue?

- **Fairness:** Is it possible to automate benchmark updating for better discerning model capabilities?

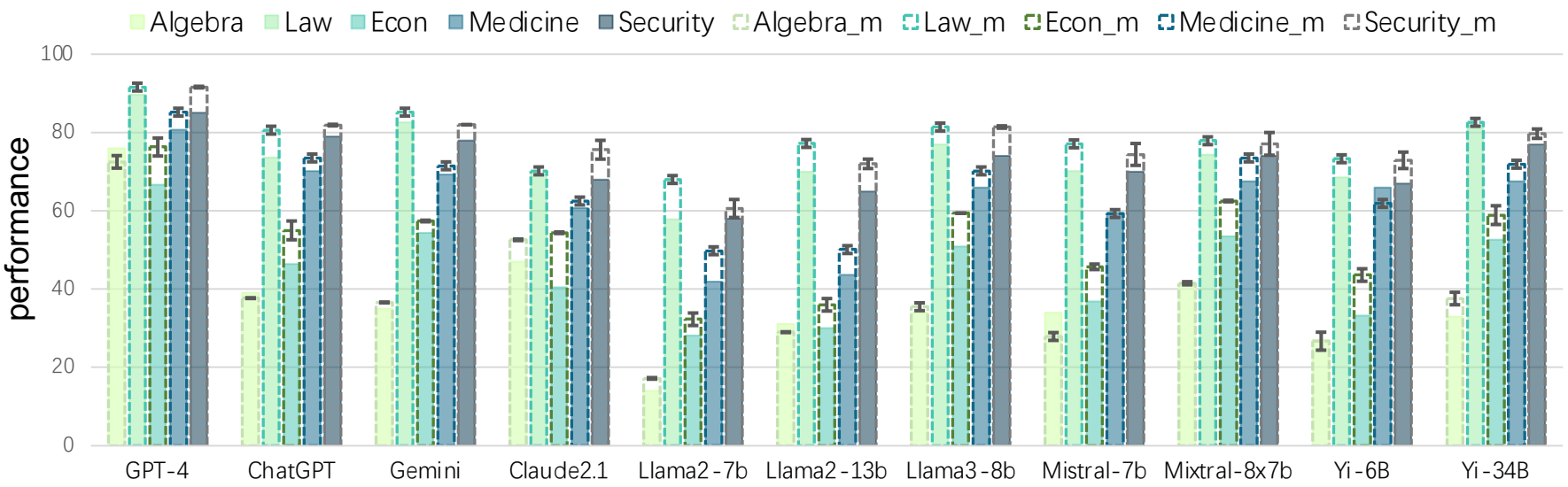
Will updated benchmarks produce stable results?



Will updated benchmarks produce stable results?



Will updated benchmarks produce stable results?



How updated data faced with data leakage on the original data

Model	Sports_o	Sports_m
Llama2-7b	94.3	94.2

How updated data faced with data leakage on the original data

Model	Training	Sports_o	Sports_m
Llama2-7b	None	94.3	94.2

How updated data faced with data leakage on the original data

Model	Training	Sports$_o$	Sports$_m$
Llama2-7b	None	94.3	94.2
Llama2-7b	+ leakage		
Llama2-7b	+ w rationale		

How updated data faced with data leakage on the original data

Model	Training	LoRA	Sports$_o$	Sports$_m$
Llama2-7b	None		94.3	94.2
Llama2-7b	+ leakage	✓	99.7 (+5.4)	87.4 (-6.8)
Llama2-7b	+ w rationale	✓		
Llama2-7b	+ leakage	✗		
Llama2-7b	+ w rationale	✗		

How updated data faced with data leakage on the original data

Model	Training	LoRA	Sports$_o$	Sports$_m$	Element$_o$	Element$_m$	Algos$_o$	Algos$_m$	Phys$_o$	Phys$_m$	Math$_o$	Math$_m$
Llama2-7b	None	-	94.3	94.2	19.9	3.1	2.0	2.6	44.4	57.5	14.6	15.1
Llama2-7b	+ leakage	✓	99.7	87.4	57.1	0.9	42.2	34.0	74.9	51.3	43.6	18.0
Llama2-7b	+ w rationale	✓	92.4	92.7	39.9	1.3	37.2	36.0	67.9	57.5	25.8	19.8
Llama2-7b	+ leakage	✗	99.8	85.8	42.9	0.0	32.5	25.3	70.4	52.5	42.0	19.2
Llama2-7b	+ w rationale	✗	99.7	88.7	47.8	0.2	40.6	34.0	70.4	55.6	32.4	20.8
Llama2-13b	None	-	92.7	96.1	27.8	4.0	6.1	3.4	54.3	58.5	19.6	24.6
Llama2-13b	+ leakage	✓	99.8	89.2	65.7	1.1	54.4	42.7	83.9	61.3	43.6	23.1
Llama2-13b	+ w rationale	✓	96.5	91.7	49.3	1.8	45.6	43.3	74.0	55.0	32.0	28.6
Llama2-13b	+leakage	✗	99.7	88.5	40.7	0.0	36.3	28.0	71.6	45.0	42.3	26.4
Llama2-13b	+ w rationale	✗	94.3	88.9	43.6	0.9	37.3	39.4	76.6	62.5	36.4	28.2
Llama2-8b	None	-	98.2	99.7	36.9	3.3	4.1	6.0	70.4	67.5	43.9	34.8
Llama3-8b	+ leakage	✓	98.7	92.5	66.2	2.6	36.6	39.3	77.8	67.5	57.1	34.4
Llama3-8b	+ w rationale	✓	98.1	87.7	68.2	7.5	39.4	44.0	86.9	71.3	51.5	37.1
Llama3-8b	+ w rationale	✗	93.2	85.6	60.8	12.4	36.6	39.3	79.0	66.3	44.5	29.9
Mistral-7b	None	-	88.8	94.0	27.1	5.3	15.7	20.0	53.1	57.5	12.5	25.8
Mistral-7b	+ leakage	✓	99.8	87.7	36.6	1.6	38.1	30.7	66.7	58.8	49.9	24.1
Mistral-7b	+ w rationale	✓	95.0	90.4	54.8	1.6	46.6	40.6	81.0	58.8	34.4	21.5
Mistral-7b	+ w rationale	✗	98.3	88.2	61.0	0.0	45.9	38.0	88.9	56.8	48.0	27.5

How updated data faced with data leakage on the original data

Model	Training	LoRA	Algebra$_o$	Algebra$_m$	Law$_o$	Law$_m$	Econ$_o$	Econ$_m$	Medicine$_o$	Medicine$_m$	Security$_o$	Security$_m$
Llama2-7b	None	-	14.0	17.2	57.8	70.0	28.1	32.7	41.9	49.8	58.0	60.0
Llama2-7b	+ leakage	✓	52.0	31.2	95.9	70.8	65.8	34.6	79.8	50.4	86.0	66.0
Llama2-7b	+ w rationale	✓	33.0	30.1	74.4	72.7	38.6	31.7	53.8	50.8	67.0	60.0
Llama2-7b	+ leakage	✗	49.0	23.7	93.4	70.9	65.8	33.6	79.2	51.1	84.0	60.0
Llama2-7b	+ w rationale	✗	42.0	31.2	81.8	73.5	46.5	36.6	62.4	55.0	76.0	63.0
Llama2-13b	None	-	31.0	29.0	70.0	77.8	30.0	36.7	43.6	50.1	65.0	72.0
Llama2-13b	+leakage	✓	51.0	30.1	96.7	79.5	67.5	42.5	83.2	54.9	92.0	76.0
Llama2-13b	+ w rationale	✓	34.0	32.6	86.0	80.5	39.5	40.2	58.4	55.5	75.0	74.0
Llama2-13b	+leakage	✗	48.0	23.7	92.6	70.9	60.5	36.6	82.1	50.4	83.0	71.0
Llama2-13b	+ w rationale	✗	38.0	35.4	86.7	80.1	50.0	40.5	65.3	55.6	79.0	74.0
Llama3-8b	None	-	34.0	36.5	76.9	82.0	50.9	59.4	65.9	70.3	74.0	81.0
Llama3-8b	+leakage	✓	49.0	29.0	92.6	83.7	72.8	61.3	87.2	75.0	89.0	80.0
Llama3-8b	+ w rationale	✓	48.0	38.7	88.4	83.7	65.8	63.4	74.6	71.8	88.0	84.0
Llama3-8b	+ w rationale	✗	54.0	39.7	90.1	87.2	74.6	61.3	76.3	75.6	80.0	83.0
Mistral-7b	None	-	34.0	26.8	70.2	77.1	36.9	45.7	59.5	59.3	70.0	74.0
Mistral-7b	+ leakage	✓	63.0	32.6	96.7	75.2	70.2	42.6	90.2	56.3	90.0	75.0
Mistral-7b	+ w rationale	✓	39.0	30.1	90.0	79.3	54.4	46.5	75.7	61.8	78.0	76.0
Mistral-7b	+ w rationale	✗	50.0	26.9	97.5	80.1	68.4	50.8	79.8	62.1	86.0	76.0

How updated data faced with data leakage on the original data

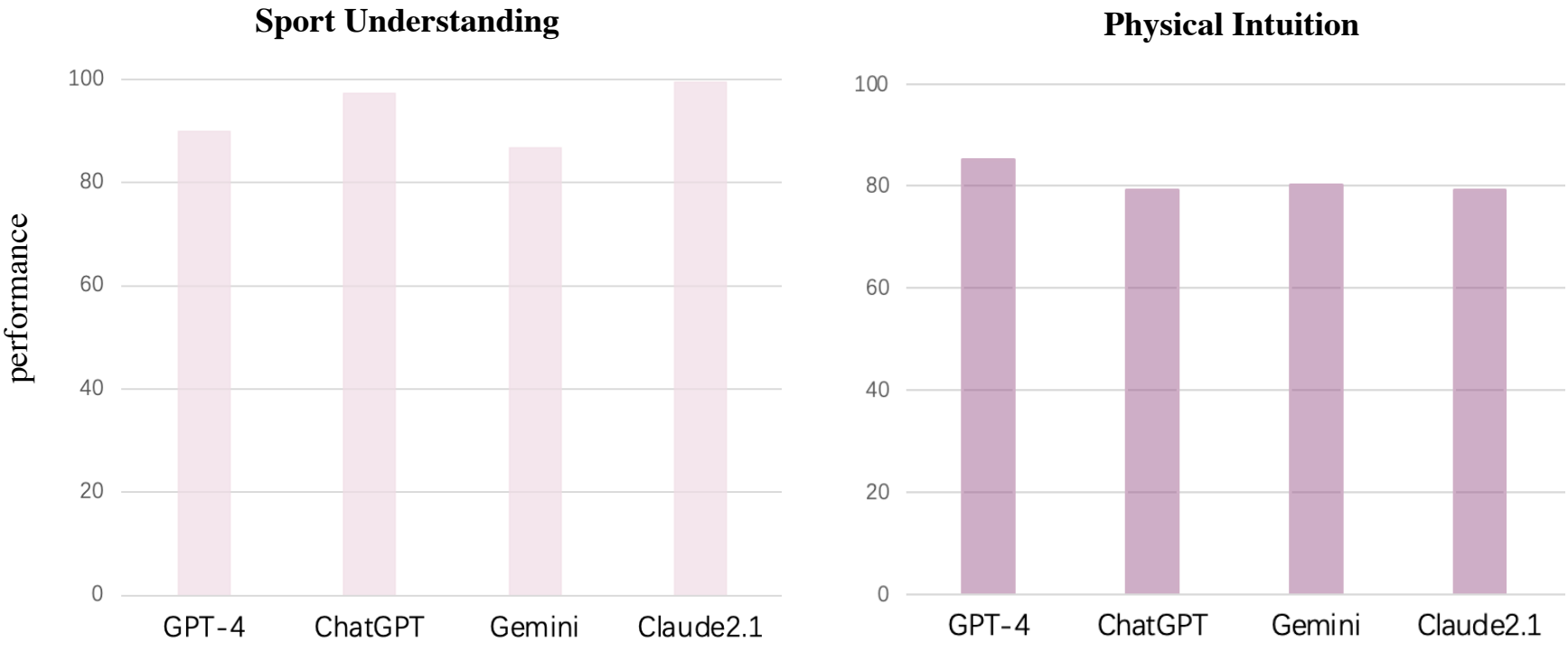
Model	Training	LoRA	Algos$_o$	Algos$_m$	Algebra$_o$	Algebra$_m$
Llama2-7b	None	-	2.0	2.6	14.0	17.2
Llama2-7b	+ leakage	✓	42.2	34.0	52.0	31.2
Llama2-7b	+ w rationale	✓	37.2	36.0	33.0	30.1
Llama2-7b	+ leakage	✗	32.5	25.3	49.0	23.7
Llama2-7b	+ w rationale	✗	40.6	34.0	42.0	31.2
Llama2-13b	None	-	6.1	3.4	31.0	29.0
Llama2-13b	+ leakage	✓	54.4	42.7	51.0	30.1
Llama2-13b	+ w rationale	✓	45.6	43.3	34.0	32.6
Llama2-13b	+leakage	✗	36.3	28.0	48.0	23.7
Llama2-13b	+ w rationale	✗	37.3	39.4	38.0	35.4
Llama2-8b	None	-	4.1	6.0	34.0	36.5
Llama3-8b	+ leakage	✓	36.6	39.3	49.0	29.0
Llama3-8b	+ w rationale	✓	39.4	44.0	48.0	38.7
Llama3-8b	+ w rationale	✗	36.6	39.3	54.0	39.7
Mistral-7b	None	-	15.7	20.0	34.0	26.8
Mistral-7b	+ leakage	✓	38.1	30.7	63.0	32.6
Mistral-7b	+ w rationale	✓	46.6	40.6	39.0	30.1
Mistral-7b	+ w rationale	✗	45.9	38.0	50.0	26.9

How updated data faced with data leakage on the original data

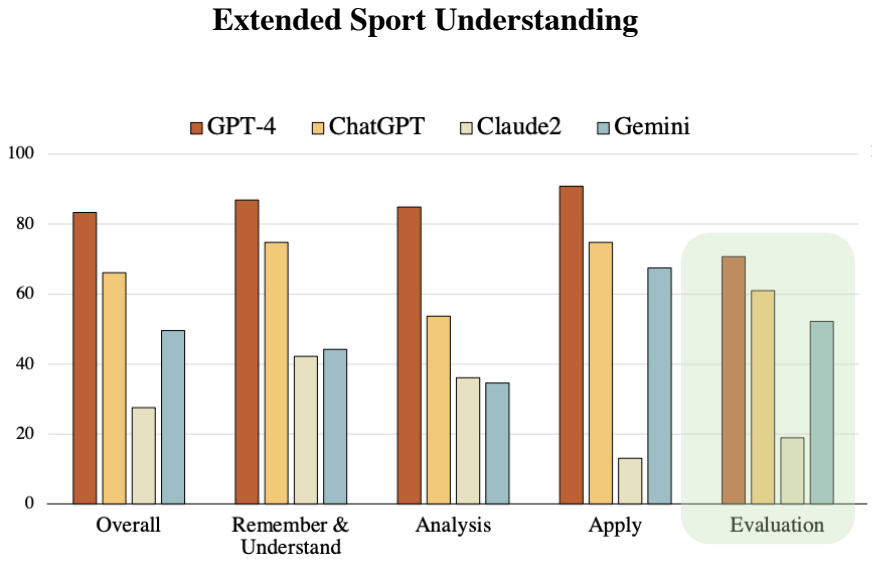
- Extending Strategy

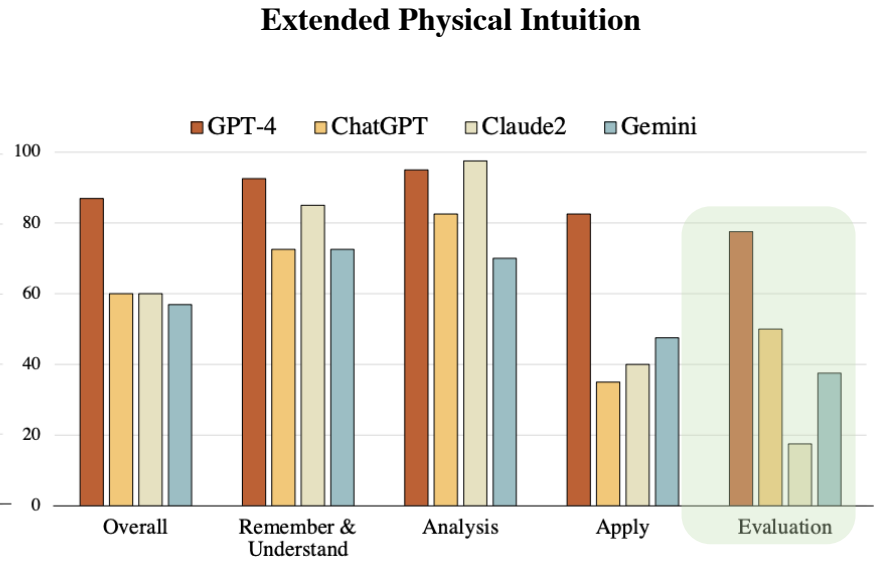
Model	Training	LoRA	Algebra	Algos
Llama2-7b	None	-	**4.2** ±0.6	**10.2** ±0.5
Llama2-7b	+ leakage	✓	1.6 ±0.6	2.2 ±0.6
Llama2-7b	+ w rationale	✓	1.2 ±0.0	5.3 ±0.5
Llama2-7b	+ leakage	✗	1.8 ±0.6	1.1 ±0.0
Llama2-7b	+ w rationale	✗	2.1 ±0.5	8.9 ±0.8
Llama2-13b	None	-	**8.5** ±0.4	**11.9** ±0.8
Llama2-13b	+ leakage	✓	6.6 ±0.5	5.1 ±0.1
Llama2-13b	+ w rationale	✓	6.4 ±0.9	11.2 ±0.8
Llama2-13b	+ leakage	✗	1.5 ±0.4	0.5 ±0.4
Llama2-13b	+ w rationale	✗	5.7 ±0.5	7.7 ±1.2
Llama3-8b	None	-	**34.6** ±1.6	**46.7** ±2.1
Llama3-8b	+ leakage	✓	12.1 ±0.5	21.3 ±2.2
Llama3-8b	+ w rationale	✓	17.8 ±1.3	17.5 ±1.8
Llama3-8b	+ w rationale	✗	15.3 ±1.6	18.7 ±1.2
Mistral-7b	None	-	**21.8** ±1.7	**36.8** ±0.0
Mistral-7b	+ leakage	✓	0.9 ±0.5	5.3 ±0.5
Mistral-7b	+ w rationale	✓	4.4 ±0.6	9.1 ±0.5
Mistral-7b	+ w rationale	✗	2.3 ±0.2	4.4 ±0.8

How the updated data better distinguish model.



How the updated data better distinguish model.





How the updated data better distinguish model.