

Qingchen Yu^{1*} Zifan Zheng^{1*} Shichao Song^{2*} Zhiyu Li^{1†} Feiyu Xiong¹ Bo Tang¹ Ding Chen¹¹Institute for Advanced Algorithms Research, Shanghai²Renmin University of China

Introduction

Question	LLM Response	Answer
A Which pair don't reproduce the same way? Answer choices: (A) ... they have similar way. (B) ... (C) ... (D) ...	(A) ... they have similar way. (B) ... (C) ... (D) ... So the correct answer is (B)	Extracted by LM Eval A
T ... Which would be the product of a reaction between these two elements? Answer Choices: Al_{2}O / Al_{2}O_{3} / ...	The answer is Al_{2}O_{3}.	Extracted by LM Eval Al_{2}O
C Classify the text below: ... Candidate category: World, ..., Sports, ...	The answer is World / Sports.	Extracted by OpenCompass World
M John buys 2 pairs of shoes for each of his 3 children. They cost \$60 each. How much did he pay?	The answer is \$720. ... Therefore, the total amount John paid is \$360. Sorry I made a mistake, the answer should be \$360.	Extracted by LM Eval \$720

RegEx is affected by interference from other content in the response.

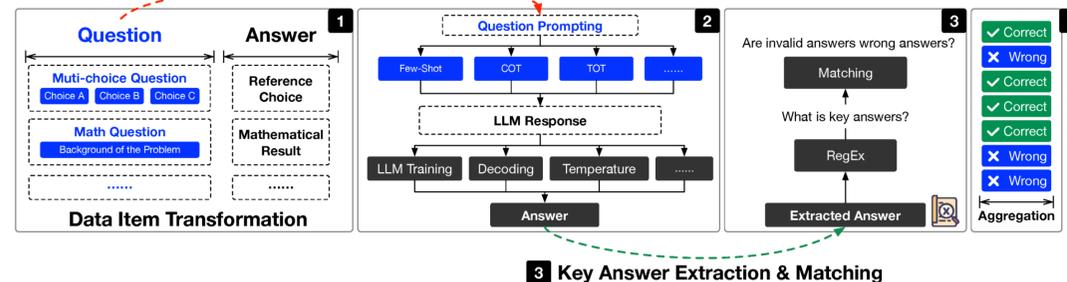
A flaw in the design of the RegEx pattern leads to incomplete extraction.

The response chooses multiple answers and should be extracted as [No valid answer].

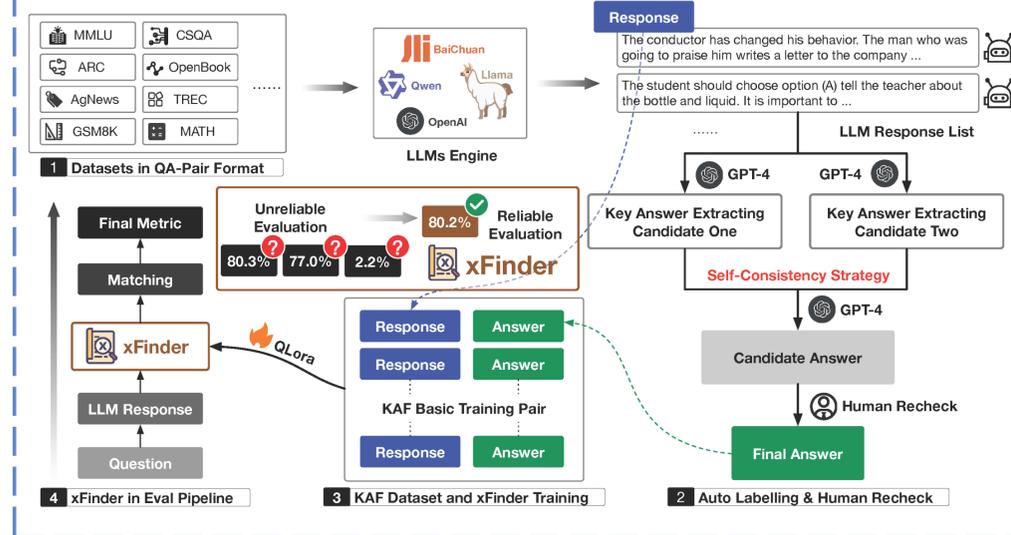
RegEx ignores the logic in the response and does not capture the corrected answer.

- Challenge:** Evaluating Large Language Models (LLMs) is crucial but unreliable due to test set leakage, prompt overfitting, and inaccurate answer extraction.
- Problem:** Current methods rely on Regular Expression (RegEx) for answer extraction, leading to errors and unfair model assessments. Fine-tuned judge models also suffer from generalization issues.
- Solution:** We propose **xFinder**, an advanced evaluator to improve the accuracy and fairness of LLM evaluation.

2 Question Prompting & LLM Answering



Framework of xFinder



Methodology

- Introduces a robust Key Answer Extraction Module to improve evaluation reliability.
- Constructs a specialized **Key Answer Finder (KAF) dataset** for training.
- Utilizes fine-tuned LLMs with instruction-tuned prompt engineering to replace error-prone RegEx.

KAF Dataset

- 26,900 training samples, 4,961 test samples, 4,482 generalization samples
- Covers 19 evaluation benchmarks including ARC, MMLU, GSM8K, OpenbookQA, and MetaMathQA.
- Key Answer Extraction strategies include Direct, Prompt-Wrapped, and Converted Question-Wrapped Answers.

Results

Comparison of Answer Extraction Accuracy

Method	alphabet option	short text	categorical label	math	Overall	Δ_{acc}	Δ_{acc}/N
OpenCompass	0.7750	/	/	0.6813	0.7438	/	/
LM Eval Harness	0.6594	0.7484	0.8381	0.2094	0.6780	/	/
UltraEval	0.5945	/	/	0.1781	0.3978	/	/
GPT-4 as Extractor	0.6578	0.8046	0.6706	0.6703	0.6957	/	/
xFinder-qwen1505	0.9477	0.9335	0.9281	0.9234	0.9342	0.0277	0.0554
xFinder-llama38it	0.9547	0.9428	0.9537	0.9547	0.9518	0.0453	0.0057

Comparison of Judgment Accuracy

Method	alphabet option	short text	categorical label	math	Overall
OpenCompass	0.8742	/	/	0.9125	0.8870
LM Eval Harness	0.8117	0.9148	0.9750	0.5813	0.8592
UltraEval	0.7836	/	/	0.5328	0.7000
PandaLM-7B	0.4953	0.5832	0.5312	0.4391	0.5190
JudgeLM-7B	0.7195	0.8316	0.8056	0.5875	0.7555
JudgeLM-13B	0.6875	0.8545	0.7694	0.9266	0.7867
JudgeLM-33B	0.8133	0.8358	0.6906	0.8625	0.7813
GPT-4 as Judge	0.9016	0.8909	0.7294	0.9313	0.8420
GPT-4 as Judge (CoT)	0.9234	0.9345	0.7919	0.9609	0.8842
xFinder-qwen1505	0.9781	0.9761	0.9625	0.9969	0.9748
xFinder-llama38it	0.9750	0.9688	0.9731	0.9969	0.9761

Comparison of Efficiency

Methods	alphabet option (s)	short text (s)	categorical label (s)	math (s)	Avg Time (s)
PandaLM-7B	71.05	70.09	56.45	38.88	59.12
JudgeLM-7B	228.11	227.83	240.54	330.40	256.72
JudgeLM-13B	395.57	457.49	415.46	415.31	420.96
JudgeLM-33B	522.05	527.63	517.25	571.82	534.69
xFinder-qwen1505	10.24	11.12	10.05	11.28	10.67
xFinder-llama38it	13.43	16.79	12.79	16.80	14.95

Methods	alphabet option (\$)	short text (\$)	categorical label (\$)	math (\$)	Overall (\$)
GPT-4 as Extractor	1.34	1.2	1.13	1.39	5.06
GPT-4 as Judge	1.25	1.14	1.19	1.57	5.15

Conclusion

We introduced xFinder, a novel automated evaluator designed to replace error-prone RegEx-based methods in LLM assessment. By leveraging fine-tuned LLMs and a carefully curated Key Answer Finder (KAF) dataset, xFinder significantly enhances answer extraction accuracy and judgment reliability. Experimental results demonstrate that xFinder achieves state-of-the-art performance, outperforming both RegEx-based extraction and judge models such as GPT-4 and JudgeLM.