Evaluating LLMs at Detecting Errors in LLM Responses

Ryo Kamoi[1], Sarkar Snigdha Sarathi Das[1], Renze Lou[1], Jihyun Janice Ahn[1], Yilun Zhao[2], Xiaoxin Lu[1], Nan Zhang[1], Yusen Zhang[1], Ranran Haoran Zhang[1], Sujeeth Reddy Vummanthala[1], Salika Dave[1], Shaobo Qin[3], Arman Cohan[2,4], Wenpeng Yin[1], Rui Zhang[1]

[1]Penn State University, [2]Yale University, [3]Stony Brook University, [4]Allen Institute for AI

{ryokamoi, rmz5227}@psu.edu

Paper & Dataset

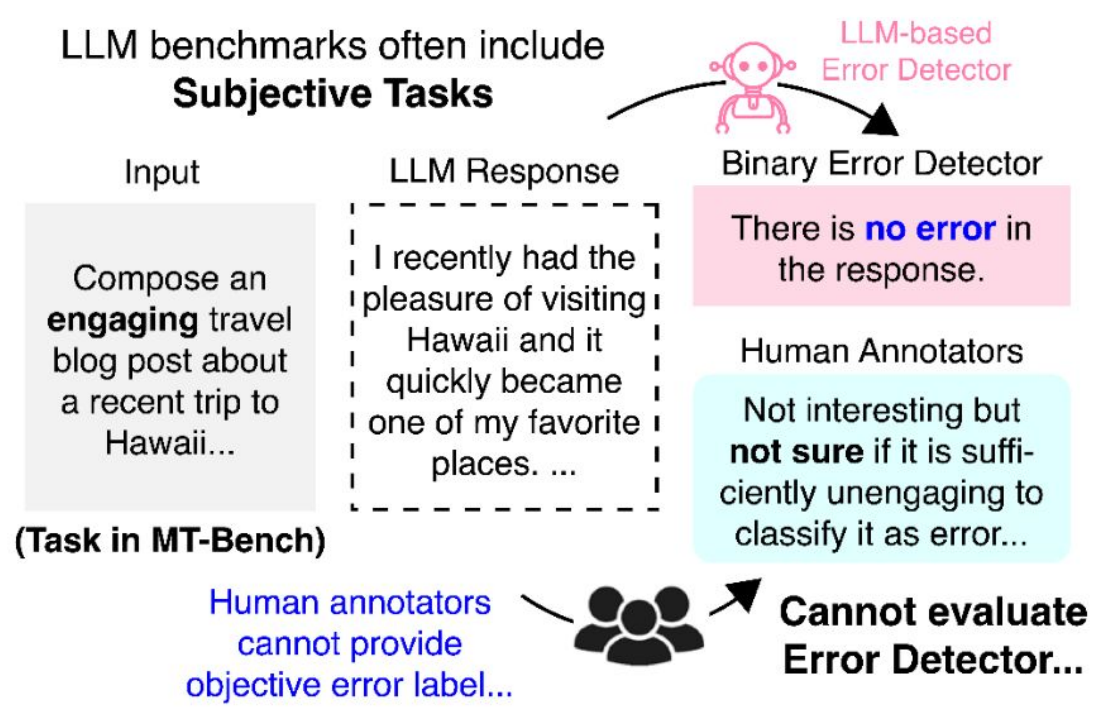
We introduce *ReaLMistake*, a benchmark for **evaluating LLMs at detecting errors in LLM responses**. Our Experiments show that even strong LLMs, such as GPT-4 and Claude 3, **detect errors made by LLMs at very low recall** and also explanations by LLM-based error detectors are unreliable.

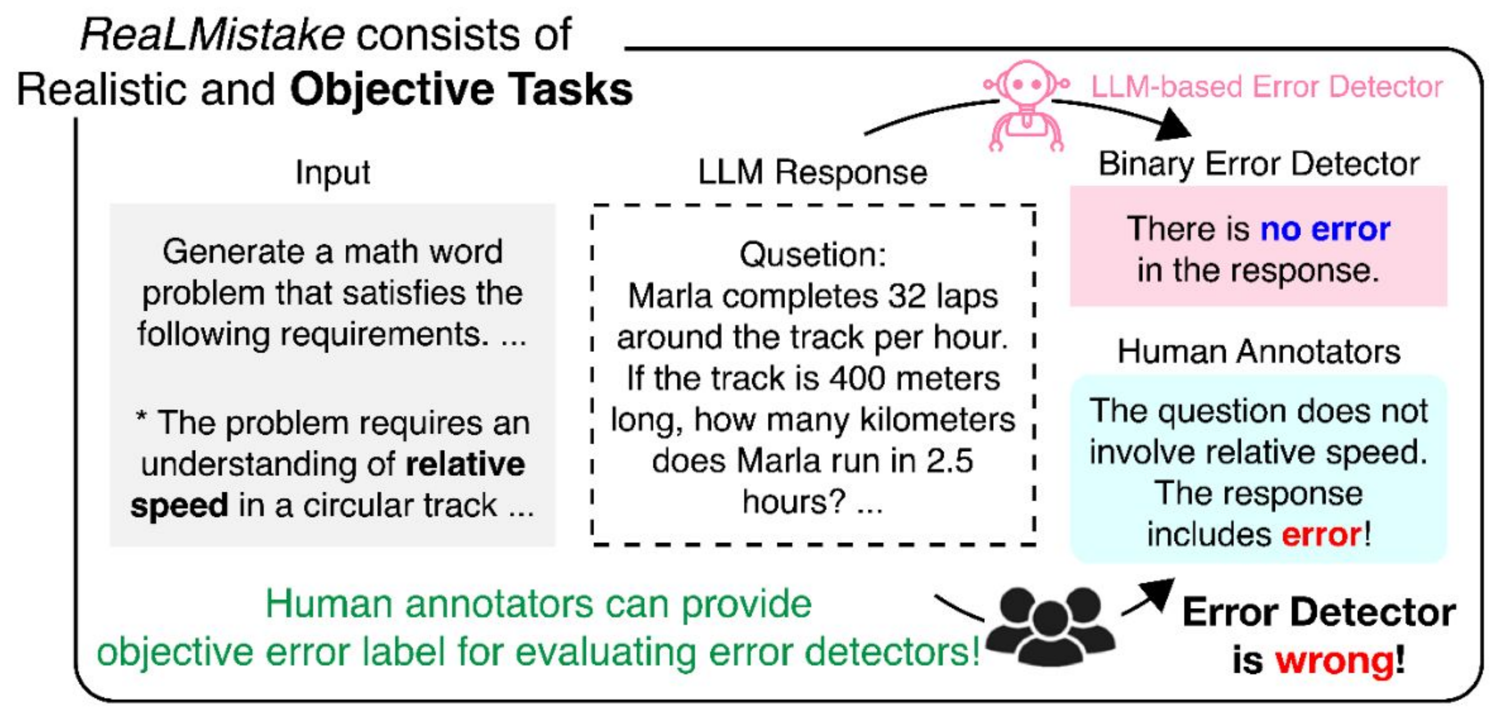
Introduction and Motivation

As LLMs have been increasingly used in real-world applications, it is critical to develop methods for automatically detecting errors in responses from LLMs. However, there is a deficiency in research specifically targeting error detection of LLM responses.

An obstacle in studying error detection is **the lack of benchmarks that include binary error annotations** (i.e., whether the response contains errors or not) on objective, realistic, and diverse errors made by LLMs.

Specifically, to provide objective error labels, tasks should not involve subjectivity or ambiguity. In many NLP tasks, even humans cannot objectively annotate binary error labels because the tasks are often open-ended and evaluation involves ambiguity.





- Many NLP tasks are ambiguous and even humans cannot annotate errors in high agreement
- We introduce the *ReaLMistake* benchmark with **realistic, objective, and diverse errors** made by LLMs (GPT-4 and Llama 2 70B) for evaluating error detection.

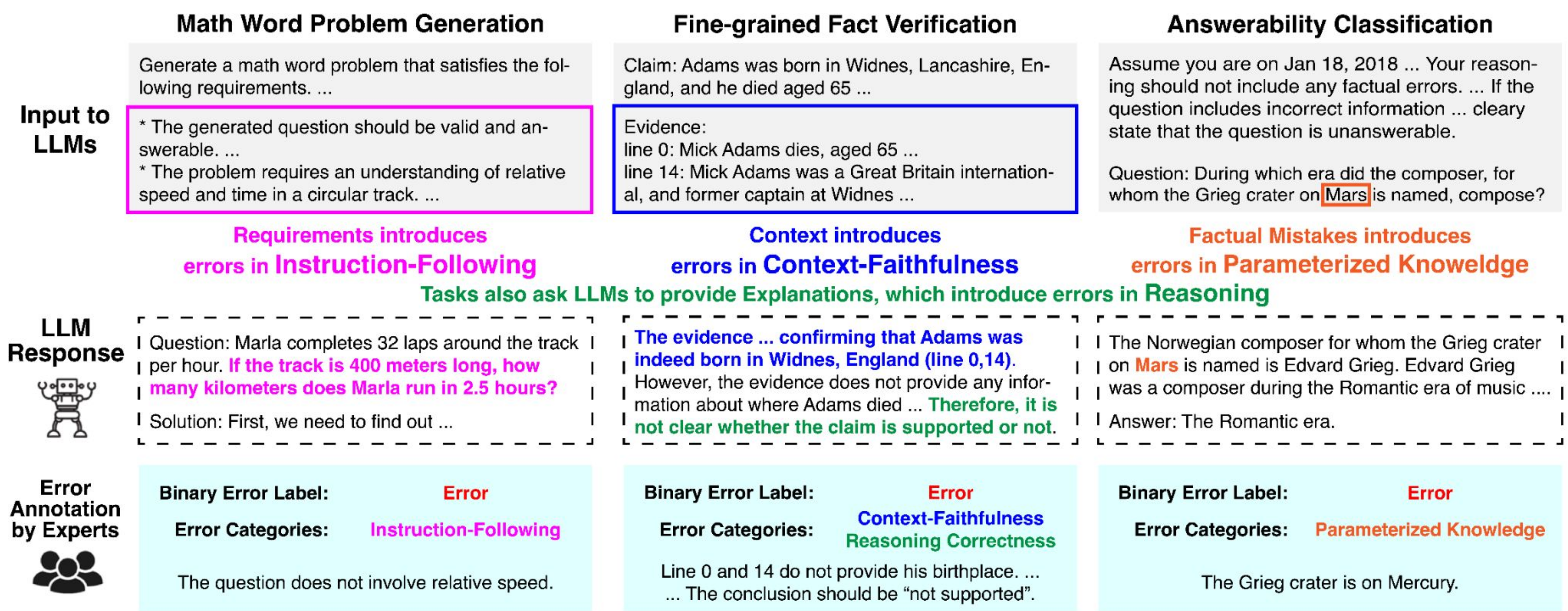
Dataset Creation

To create tasks that satisfy the requirements, we propose an approach to **design tasks so that they make LLMs introduce errors detected by objective, realistic, and diverse evaluation criteria**. We identify four criteria that can be objectively evaluated by humans and cover diverse errors in LLM responses:

- Instruction-Following
- Context-Faithfulness
- Parameterized Knowledge
- Reasoning

We create three tasks with the intention of making LLMs introduce errors detected by these four evaluation criteria, **eliminating subjectivity from the error annotation process**.

We create the ReaLMistake dataset by collecting error annotations on 900 responses from GPT-4 and Llama 2 70B in the three tasks. The annotation process requires careful checking of the entire LLM responses, and 14 expert annotators spent 90 hours in total to provide high-quality annotations.



Dataset Creation Process of ReaLMistake

- We identify four categories of errors that can be objectively evaluated by humans
- We design three tasks so that LLM responses only include these four objective errors

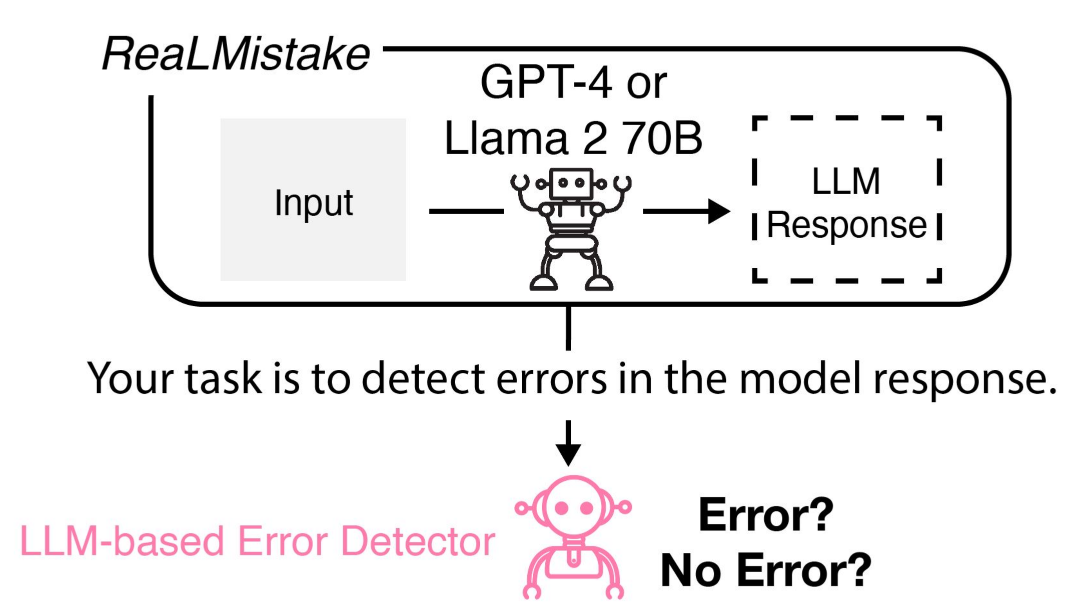
Experiments

We evaluate LLM-based error detectors with zero-shot prompts using 12 LLMs:

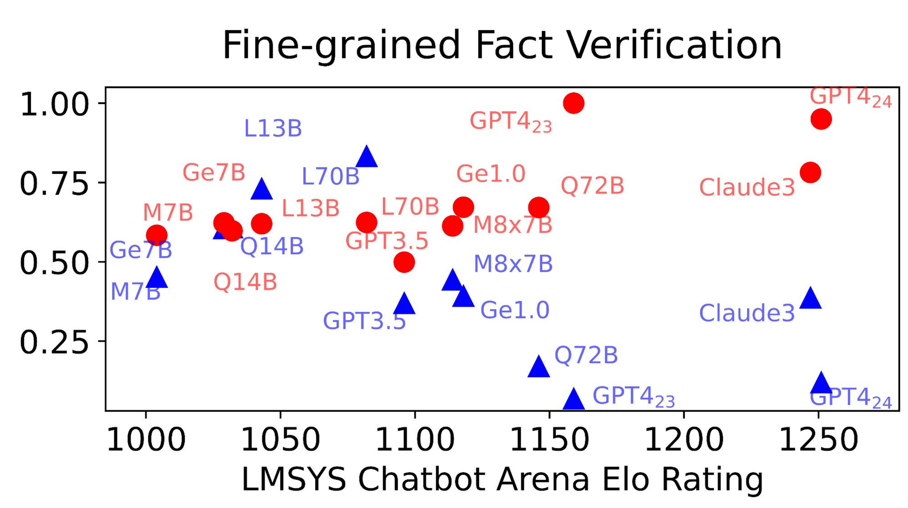
- Zero-shot chain-of-thought (4 prompts)
- Self-consistency
- Majority vote by multiple LLMs
- G-Eval style human-written instruction

However, **all detectors are much worse than human performance and are often even worse than random baselines**.

In addition, our manual analysis shows that explanations generated by LLM-based error detectors are often wrong, even when the final answer (error or no error) is corect.



Error Detection Task in ReaLMistake



Precisions and recalls of 12 LLM-based error detectors on ReaLMistake (errors in GPT-4 responses)

- Better LLMs achieve better precision (red circles) but with **lower recall** (blue triangles)
- Strong LLMs are conservative about detecting mistakes and **miss many errors in LLM responses**!

		Gemma	Llama 2		Mistral		Qwen 1.5		GPT-3.5	Gemini	Claude 3	GPT-4		Random	Human
		7B	13B	70B	7B	8x7B	14B	72B	0125	1.0 Pro	Opus	0613	0125		
GPT-4 0613	MathGen	46.5	54.2	59.5	6.9	45.5	52.3	32.8	65.3	42.5	50.1	63.1	**70.9**	62.1	90.0
	FgFactV	60.3	65.4	**69.9**	50.9	46.8	57.7	24.9	41.4	45.8	48.9	12.7	20.8	62.9	95.5
	AnsCls	59.2	69.8	**69.8**	48.1	38.3	53.8	15.1	28.8	40.7	38.5	20.0	22.1	62.1	90.5
Llama 2 70B	MathGen	54.3	56.6	69.2	9.0	56.0	54.9	50.3	72.3	52.9	81.8	88.7	**90.8**	80.0	98.3
	FgFactV	68.9	78.7	**81.8**	68.2	35.1	64.6	18.3	34.2	42.0	45.2	38.8	68.5	80.6	100.0
	AnsCls	34.8	**77.4**	51.6	61.9	29.8	44.9	5.1	3.7	16.4	23.2	61.6	75.9	81.2	100.0

F1 scores of 12 LLM-based error detectors (zero-shot CoT) on ReaLMistake. Gray color represent values **worse than the random baseline**.