Evaluating LLMs at Detecting Errors in LLM Responses



Ryo Kamoi¹, Sarkar Snigdha Sarathi Das¹, Renze Lou¹, Jihyun Janice Ahn¹, Yilun Zhao², Xiaoxin Lu¹, Nan Zhang¹, Yusen Zhang¹, Ranran Haoran Zhang¹, Sujeeth Reddy Vummanthala¹, Salika Dave¹, Shaobo Qin³, Arman Cohan^{2,4}, Wenpeng Yin¹, Rui Zhang¹

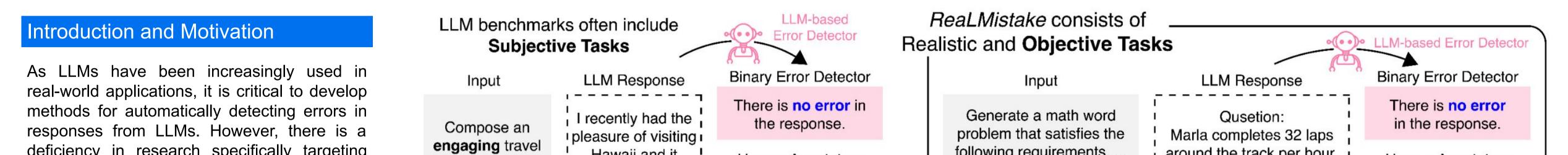


¹Penn State University, ²Yale University, ³Stony Brook University, ⁴Allen Institute for Al {ryokamoi, rmz5227}@psu.edu



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.



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.

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:

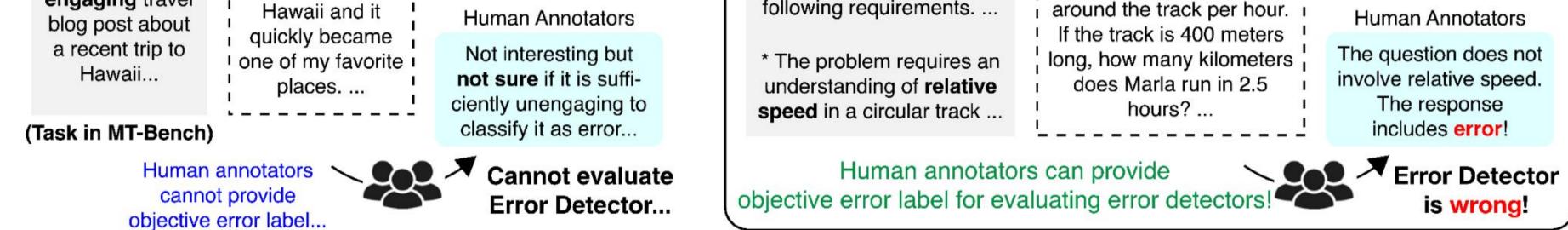
• Instruction-Following

Dataset Creation

- 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.



- 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.

	Math Word Problem Generation		Fine-grained	Fact Verification		Answerability Classification							
Input to LLMs	Generate a math word problem that satisfies the fol- lowing requirements		Claim: Adams was born gland, and he died aged	in Widnes, Lancashire, En- 65		Assume you are on Jan 18, 2018 Your reason- ing should not include any factual errors If the							
	* The generated question should be valid and an- swerable		Evidence: line 0: Mick Adams dies,	•	question includes incorrect information cleary state that the question is unanswerable.								
	* The problem requires an understanding of relative speed and time in a circular track		al, and former captain at	a Great Britain internation- Widnes		Question: During which era did the composer, for whom the Grieg crater on Mars is named, compose?							
	Requirements introduces errors in Instruction-Following			introduces		Factual Mistakes introduces errors in Parameterized Knoweldge							
	errors in Instruction-Following errors in Context-Faithfulness errors in Parameterized Knoweldge Tasks also ask LLMs to provide Explanations, which introduce errors in Reasoning												
LLM Response	 I Question: Marla completes 32 laps around the track I per hour. If the track is 400 meters long, how I many kilometers does Marla run in 2.5 hours? I Solution: First, we need to find out 	- 	indeed born in Widne However, the evidence mation about where Ad	rming that Adams was s, England (line 0,14). does not provide any infor- ams died Therefore, it is claim is supported or not.	1	I The Norwegian composer for whom the Grieg crater I I on Mars is named is Edvard Grieg. Edvard Grieg I was a composer during the Romantic era of music							
Error	Binary Error Label: Error		Binary Error Label:	Error		Binary Error Label:	Error						
Annotation by Experts	Error Categories: Instruction-Following		Error Categories:	Context-Faithfulness Reasoning Correctness		Error Categories:	Parameterized Knowledge						
	The question does not involve relative speed.			t provide his birthplace hould be "not supported".		The Grieg crater is on Mercury.							

Experiments

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

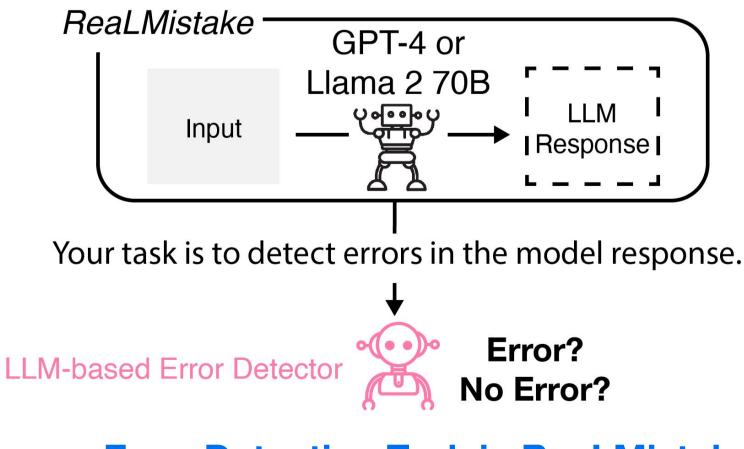
- Zero-shot chain-of-thought (4 prompts)
- Self-consistency
- Majority vote by multiple LLMs
- G-Eval style human-written instruction

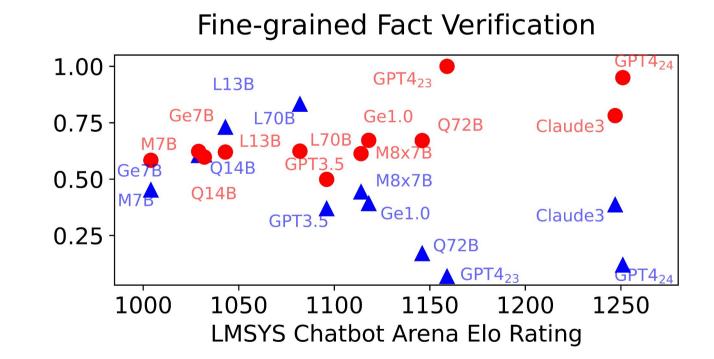
However, all detectors are much worse than human performance and are often even worse than random baselines.

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.

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





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

Error Detection Task in ReaLMistake

- 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		Dandam	Human
		7B	13B	70B	7B	8x7B	14B	72B	0125	1.0 Pro	Opus	0613	0125	Random	Human
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
	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
Llama 2	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



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