中国科学院大学
University of Chinese Academy of Sciences

Reasoning Errors of LLMs

Wanli Yang

Mar 14, 2025

STAR Group Paper Reading

Table of Contents

Overview

Overview

Paper List:

- Large Language Models Cannot Self-Correct Reasoning Yet (**ICLR24**, 370+ citations)
- LLMs cannot find reasoning errors, but can correct them given the error location (**ACL24 Findings**, 80+ citations)
- Evaluating LLMs at Detecting Errors in LLM Responses (**COLM24**, 10+ citations)

LLMs Can’t Self-Correct Reasoning

LARGE LANGUAGE MODELS CANNOT SELF-CORRECT REASONING YET

Jie Huang[1,2*] **Xinyun Chen**[1*] **Swaroop Mishra**[1] **Huaixiu Steven Zheng**[1] **Adams Wei Yu**[1] **Xinying Song**[1] **Denny Zhou**[1]

[1]Google DeepMind [2]University of Illinois at Urbana-Champaign

jeffhj@illinois.edu, {xinyunchen, dennyzhou}@google.com

Background

- Leading LLMs may still generate incorrect response
- "**Self-correction**" emerged as a promising solution
- LLMs **refine** their responses based on **feedback** to their **previous outputs**

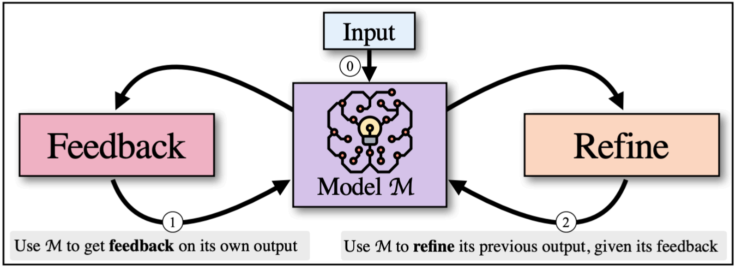


Figure from "Self-Refine: Iterative Refinement with Self-Feedback"(NIPS2023).

Motivation

- If an LLM possesses the ability to self-correct, why doesn't it simply offer the correct answer in its initial attempt?
- (LLMs know more than they express?)
- Delves into the paradox, **critically examining** the **self-correction** capabilities of LLMs on **reasoning**.

Source of Feedback

Pivotal definition distinction lies in source of feedback:

- **Internal feedback**: parametric knowledge
- **External inputs**: humans, other models, tools, and knowledge sources

This paper focuses on **intrinsic self-correction**

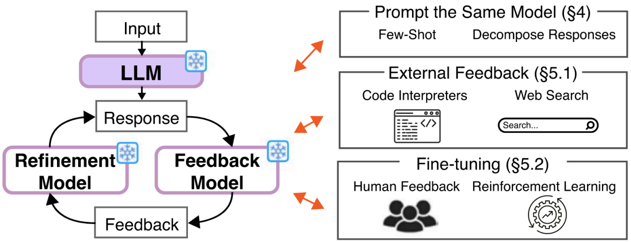


Figure from “When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs” (TACL2024)

Experimental Setup

Benchmarks:

- **GSM8K**: diverse grade school math word problems
- **CommonSenseQA**: multi-choice questions that test commonsense reasoning
- **HotpotQA**: multi-hop question answering dataset

```
"GSM8K": {"question": "Natalia sold clips to 48 of her friends in April, and then she sold
↪  half as many clips in May. How many clips did Natalia sell altogether in April and
↪  May?"}
"CommonSenseQA": {"question": "The sanctions against the school were a punishing blow, and
↪  they seemed to what the efforts the school had made to change?"}
"HotpotQA": {"question": "What was the former band of the member of Mother Love Bone who
↪  died just before the release of 'Apple'?"}
```

Experimental Setup

Test Models:

- Self-correction with **oracle labels**:
  - GPT-3.5-Turbo
  - GPT-4
- **Intrinsic** self-correction: (+)
  - GPT-4-Turbo
  - Llama-2-70b-chat

Setup:

- Prompt the models to undergo a **maximum of two rounds** of self-correction
- **Temperature of 1** for GPT-3.5-Turbo and GPT-4, and **temperature of 0** for GPT-4-Turbo and Llama-2

Experimental Setup

Prompts: apply a three-step prompting strategy for self-correction

- Prompt for an **initial generation**
- Prompt model to review and produce **feedback**
- Prompt model to **answer** with feedback

```
Can you solve the following math problem? Christina is planning a birthday party ......
↪  How much will she spend? Explain your reasoning. Your final answer should be a single
↪  numerical number, in the form \boxed{answer}, at the end of your response.

Review your previous answer and find problems with your answer.

Based on the problems you found, improve your answer. Please reiterate
your answer, with your final answer a single numerical number, in the form \boxed{answer}.
```

Results with Oracle Labels

Strategy: use **correct label** to determine **when to stop** self-correction loop

Self-correction with oracle labels showcases **significant performance improvements**

		GSM8K	CommonSenseQA	HotpotQA
GPT-3.5	Standard Prompting	75.9	75.8	26.0
	Self-Correct (Oracle)	84.3	89.7	29.0
GPT-4	Standard Prompting	95.5	82.0	49.0
	Self-Correct (Oracle)	97.5	85.5	59.0

But the availability of oracle labels seems counter-intuitive

Results of Intrinsic Self-Correction

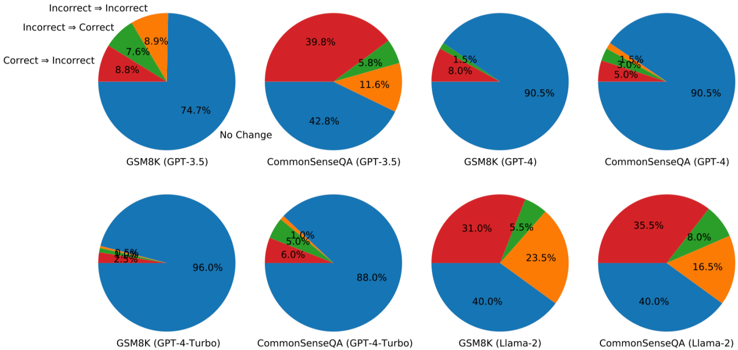
Strategy: requiring LLMs to **independently determine when to stop** the self-correction process, i.e., **whether to retain their previous answers**

The accuracies of all models drop across all benchmarks even with different prompts

		# calls	GSM8K	CommonSenseQA	HotpotQA
GPT-3.5	Standard Prompting	1	**75.9**	**75.8**	**26.0**
	Self-Correct (round 1)	3	75.1	38.1	25.0
	Self-Correct (round 2)	5	74.7	41.8	25.0
GPT-4	Standard Prompting	1	**95.5**	**82.0**	**49.0**
	Self-Correct (round 1)	3	91.5	79.5	**49.0**
	Self-Correct (round 2)	5	89.0	80.0	43.0

Empirical Analysis of Decline

- **More likely to modify a correct answer to incorrect** than to revise an incorrect answer
- Fundamental issue: LLMs **cannot properly judge the correctness** of their reasoning
- GPT-4 and GPT-4-Turbo are **more robust**



Intuitive Explanation of Decline

Terry eats 2 yogurts a day. They are currently on sale at 4 yogurts for $5.00. How much does he spend on yogurt over 30 days?

... So, Terry spends $\boxed{75.00}$ on yogurt over 30 days. (Correct)

Review your previous answer and find problems with your answer.

Upon reviewing my previous answer, I realized that I made a mistake in calculating Terry's spending on yogurt over 30 days. I incorrectly stated that Terry spends $2.50 per day for 30 days, resulting in a total of $75.00…

Based on the problems you found, improve your answer.

... Therefore, the final answer is Terry spends $\boxed{37.50}$ on yogurt over 30 days. (Incorrect)

- Well-aligned model paired with thoughtfully designed **initial prompt should obtain optimal response**
- **Supplementary feedback** may even **bias** the model away from producing an optimal response

Multi-Agent Debate Vs Self-Consistency

Multi-Agent debate: multiple instances of a single model **critique and debate**

Self-Consistency: a model generate multiple responses and performs **majority voting**

Equivalent number of responses: multi-agent debate **significantly underperforms** self-consistency

	# responses	GSM8K
Standard Prompting	1	76.7
Self-Consistency	3	82.5
Multi-Agent Debate (round 1)	6	83.2
Self-Consistency	6	85.3
Multi-Agent Debate (round 2)	9	83.0
Self-Consistency	9	**88.2**

Prompt Design Issues in Self-Correction

Motivation: **Initial** prompt should be **informative enough** for **fair comparison**

Previous work [1] does **not** clearly **specify all** the **requirements** in initial prompt

- Optimizing initial prompt significantly **outperforms** self-correction
- Self-correction on optimized prompts **leads to decreased performance**

	# calls	CommonGen-Hard
Standard Prompting*	1	44.0*
Self-Correct*	7	67.0*
Standard Prompting*	1	53.0
Self-Correct*	7	61.1
Standard Prompting (ours)	1	**81.8**
Self-Correct*	7	75.1

* Prompts and results from Madaan et al. (2023).

Conclusion

LLMs cannot self-correct reasoning yet:

- Existing achievements rely on oracle labels;
- Multi-agent debate for self-correction underperforms self-consistency;
- Uninformative initial prompts lead to inflated results.

LLMs Can’t Find Errors,
but Can Correct with Location

LLMs cannot *find* reasoning errors, but can *correct* them given the error location

Gladys Tyen*[1], Hassan Mansoor[2], Victor Cărbune[2], Peter Chen†[2], Tony Mak†[2]
[1]University of Cambridge, Dept. of Computer Science & Technology, ALTA Institute
[2]Google Research
gladys.tyen@cl.cam.ac.uk
{vcarbune,hassan,chenfeif,tonymak}@google.com

Motivation

- Previous work present self-correction as **a single process**
- This paper divides it into **mistake finding** and **output correction** to better understand each component individually

BIG-Bench Mistake

Benchmark Construction Process:

- Select **5 tasks from** the **BIG-bench** dataset: word sorting, tracking shuffled objects, logical deduction, multi-step arithmetic, and Dyck languages
- Generate **CoT-style traces** with PaLM 2 Unicorn (temperature = 0)
- Sample 255 instances with **incorrect** answer and 45 with **correct** answer for each task
- Annotate the location of **the first logical error** (majority of at least 3 annotators)

BIG-Bench Mistake Example

Sort the following words alphabetically: List: hypochlorite ponderosa phone credulity

Thought 1:	I should start by looking at the first letter of the words in the list. The first letter: "hypochlorite": "h" (8). "ponderosa": "p" (16). "phone": "p" (16). "credulity": "c" (3).
Thought 2:	We now have: (3) "credulity" < (8) "hypochlorite" < (16) ["ponderosa" ? "phone"].
Thought 3:	Now let's sort this subpart ["ponderosa" ? "phone"] by looking at their second letters. The second letter: "ponderosa": "o" (15). "phone": "h" (8).
Thought 4: **(MISTAKE)**	We now have: (8) "phone" < (15) "ponderosa" for the subpart. Hence, we have "credulity" < "phone" < "ponderosa".
Thought 5:	I have now sorted all the words. The answer is credulity hypochlorite phone ponderosa

Can LLMs Find Reasoning Mistakes?

Experimental Setup:

- **Test Models**: GPT-4-Turbo, GPT-4, GPT-3.5-Turbo, Gemini Pro, and PaLM 2 Unicorn
- **Requirements**: location matches exactly, or output correctly indicates no mistakes
- **Prompting Strategies**: 3-shot augmentation
  - Direct trace-level prompting
  - Direct step-level prompting
  - CoT step-level prompting

Can LLMs Find Reasoning Mistakes?

Model	Direct (trace)	Direct (step)	CoT (step)
Word sorting (11.7)			
GPT-4-Turbo	36.33	33.00	–
GPT-4	35.00	44.33	34.00
GPT-3.5-Turbo	11.33	15.00	15.67
Gemini Pro	10.67	–	–
PaLM 2 Unicorn	11.67	16.33	14.00
Overall			
GPT-4-Turbo	30.13	48.33	–
GPT-4	39.80	52.87	43.40
GPT-3.5-Turbo	10.44	14.78	14.31
Gemini Pro	16.14	–	–
PaLM 2 Unicorn	17.09	23.67	24.65

Results:

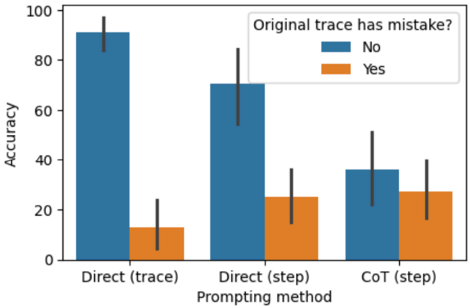
- **Direct step-level prompting GPT-4** attains **best** results but only reaches accuracy of **52.87%**
- Existing self-correction strategies are **ineffective on reasoning errors**.
- If LLMs are **unable to identify mistakes**, it should be no surprise that they are **unable to self-correct** either

Comparison of Prompting Methods

From direct trace-level prompting to CoT step-level prompting

- Accuracy on traces with **mistakes arises**
- Accuracy on traces with **no mistakes goes down**

The more calls made, the more likely the model will identify at least one mistake



Can LLMs Correct Reasoning Mistakes

Objective: Examine LLMs' ability to **self-correct** mistakes, **independently of their ability to find them**. (feed oracle mistake location)

Pipeline:

- (a) Generate an initial CoT trace using **temperature = 0**

Can LLMs Correct Reasoning Mistakes

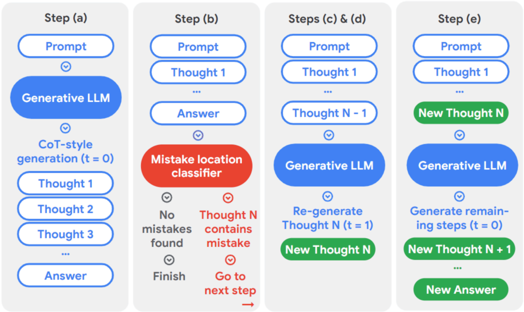
Pipeline:

- (b) Determine mistake location in this trace
- (c) **Prompt** model **again** for the same step but at **temperature = 1**
  (No mistakes, move onto next trace)

Can LLMs Correct Reasoning Mistakes

Pipeline:

- (c) often produces steps that are identical to the original
- (d) Repeat (c) **until a different step** is generated (maximum re-generation times = 8)
- (e) **Regenerated in place of previous**, then generate remaining at temperature = 0



Additional Setup

- **Comparison with Random Location**: feeding **mistake location** vs **random location** to demonstrate performance increases not from randomly resampling outputs
- Perform backtracking on both **correct**$_{ans}$ and **incorrect**$_{ans}$ traces, as long as **there is a mistake** in one of the steps

Experimental Results

- Gains from **correcting are larger** than losses from changing correct answers (Suitable for **low-accuracy** tasks)
- Random baseline improves, but are considerably smaller than mistake location
- **With mistake location available, LLMs can correct their own outputs**, suggesting **main bottleneck** of self-correction in **mistakes findings** rather than correcting

Task	With **mistake** location Δ **accuracy** ✓	With **mistake** location Δ**accuracy** ✗	With **random** location Δ **accuracy** ✓	With **random** location Δ**accuracy** ✗	Avg. num. of steps
Word sorting	-11.11	+23.53	-15.56	+11.76	11.7
Tracking shuffled objects	-6.67	+43.92	-6.67	+20.39	5.4
Logical deduction	-11.43	+36.86	-13.33	+21.57	8.3
Multistep arithmetic	-0.00	+18.04	-8.89	+10.59	5.0
Dyck languages	-6.82	+18.06	-15.91	+5.16	24.5

Obtain Mistake Location with Classifier

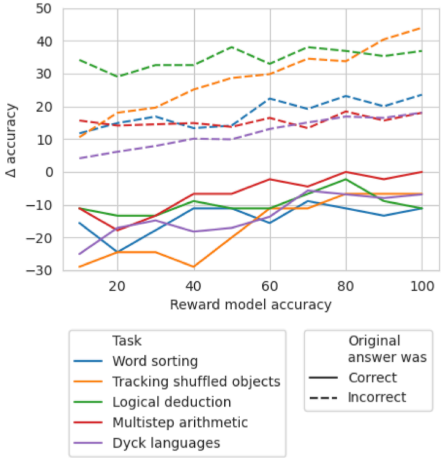
Observation:

- LLMs **fails to identify** mistake location
- LLMs **can correct** their own CoT traces with mistake location

Investigation:

obtain mistake location from a smaller, trained classifier (LLMs)

Obtain Mistake Location with Classifier



- **Question**: What mistake-finding **accuracy** is **required** to be effective?
- **Strategy**: **Simulate classifiers at different levels of accuracy** and run backtracking
- **Results**: Acc beyond 60-70% is effective

Obtain Mistake Location with Classifier

- **Question**: Is it possible to **train** a classifier **with OOD data**?
- **Strategy**: Train on 4 tasks, test on the remaining task
- **Results**: Better than self-identification, but do **not meet the required threshold**
- **Idea**: Maybe use **uncertainty**?

Held-out task	Trained classifier accuracy$_{mis}$ (Otter)	3-shot prompting accuracy$_{mis}$ (Unicorn)	Difference
Word sorting	**22.33**	11.67	+11.66
Tracking shuffled objects	**37.67**	18.00	+19.67
Logical deduction	6.00	**6.67**	-0.67
Multi-step arithmetic	**26.00**	22.00	+4.00
Dyck languages	**33.57**	10.98	+22.59

Related Work

Time of correction:

- Updating weights during **training**
- Modifying parameters during **post-training**
- Adjusting **during generation**
- Correction **on generated output**

Conclusion

- LLMs **fail to find** reasoning errors
- LLMs **can correct** them given the error location
- Train **a classifier with OOD data to find mistakes** may be effective

LLMs Detect Errors in Responses

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

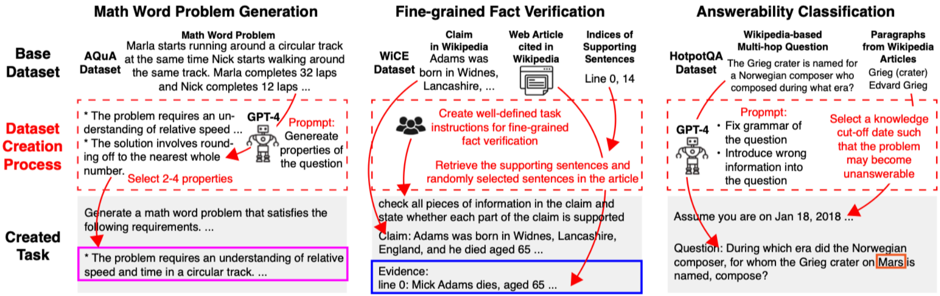
Motivation

- Systematically examine the **capabilities of LLMs in detecting response errors**
- Previous research focuses on tasks of **little practical value** (word sorting) or **limited error types** (faithfulness in summarization)
- This paper introduces **ReaLMistake**, the first error detection benchmark consisting of **objective, realistic, and diverse errors** made by LLMs

RealMistake

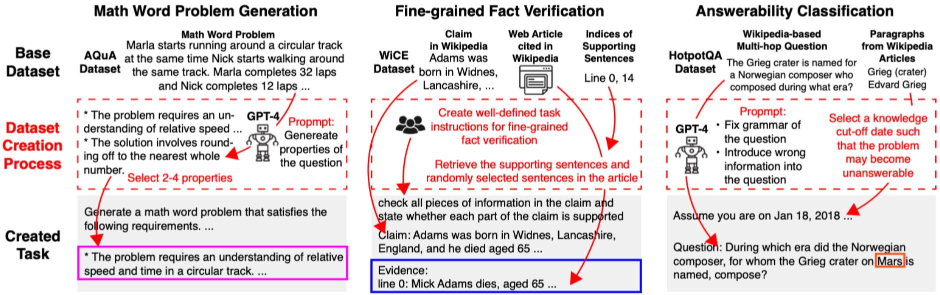
Tasks:

- Math Word Problem Generation
- Fine-grained Fact Verification
- Answerability Classification



Criteria:

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



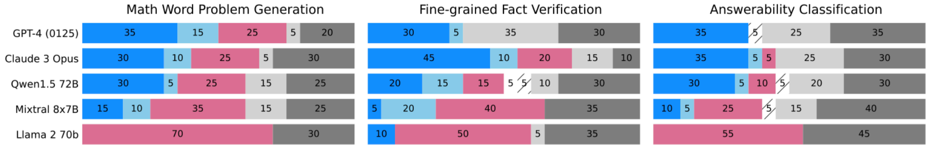
Results

Error detection task is difficult even for Claude 3 and GPT-4: high precision but low recall

Error Detector		Gemma 7B	Llama 2 13B	Llama 2 70B	Mistral 7B	Mistral 8x7B	Qwen 1.5 14B	Qwen 1.5 72B	GPT3.5 0125	Gemini 1.0 Pro	Claude3 Opus	GPT-4 0613	GPT-4 0125	Random	Expert Human
F1															
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
Precision															
GPT-4 0613	MathGen	61.6	62.6	73.0	22.8	75.5	77.4	82.9	77.3	78.1	**94.9**	94.4	88.9	62.1	100.0
	FgFactV	62.3	62.0	62.4	58.4	61.3	59.8	67.1	49.9	67.2	78.2	**100.0**	95.0	62.9	95.5
	AnsCls	64.0	62.2	65.2	59.8	60.9	68.6	55.4	72.8	78.4	74.9	79.9	**88.2**	62.1	95.0
Llama 2 70B	MathGen	82.6	79.5	88.6	41.8	89.0	96.2	94.5	86.4	90.0	95.0	**97.7**	95.2	80.0	100.0
	FgFactV	83.5	81.9	82.4	80.0	96.3	83.2	73.7	98.7	85.7	**99.3**	85.4	92.6	80.6	100.0
	AnsCls	80.5	82.5	77.3	83.8	86.3	74.8	70.5	69.4	78.3	**100.0**	97.1	98.4	81.2	100.0
Recall															
GPT-4 0613	MathGen	50.0	52.3	**75.3**	4.3	35.1	49.7	23.3	64.1	41.7	35.9	48.0	59.5	62.1	81.8
	FgFactV	60.5	73.0	**83.2**	45.2	44.3	60.8	17.0	36.9	39.2	38.6	6.8	11.9	62.9	95.5
	AnsCls	57.2	**81.3**	79.3	45.4	29.6	54.0	8.9	19.3	31.6	26.4	11.5	12.6	62.1	86.4
Llama 2 70B	MathGen	51.2	50.2	72.9	5.7	44.3	47.3	37.5	65.8	46.9	72.7	81.2	**86.9**	80.0	96.7
	FgFactV	61.8	77.5	**82.9**	60.7	24.4	61.2	11.0	24.2	32.2	32.6	25.8	54.8	80.6	100.0
	AnsCls	23.3	**77.5**	46.7	52.3	19.4	45.2	2.7	1.9	9.8	13.3	45.2	62.1	81.2	100.0

Unreliable Explanations

Explanations by **open-source models** are **more often wrong** even when the binary **predictions are correct**.

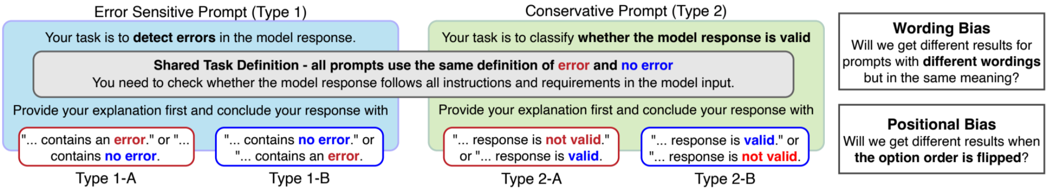


● Correct prediction & explanation ● Correct prediction & wrong explanation ● Wrong prediction & explanation

Error Detection is Sensitive to Prompt

Recall of error detection is sensitive to small changes in prompts

- **Positional Bias**: “error” option first has 16.0 ± 21.7% (Type 1) and 27.2 ± 23.9% (Type 2) higher recall
- **Wording Bias**: In an average of 12 LLMs and 3 tasks, Type 1 (error) has 16.9 ± 20.3% higher recall



Brainstorm

Brainstorm

How can we avoid mistakes in LLMs Reasoning?

- Practicality of **correction on generated contents** (compared to correction during generation?)
- **Uncertainty** to avoid mistakes in reasoning? (Low uncertainty then RAG)
- RAG is effective for factual errors, but what about **logical error**?
- ……

Thanks for Listening!

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark.
Self-refine: Iterative refinement with self-feedback.
In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc., 2023.