When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs

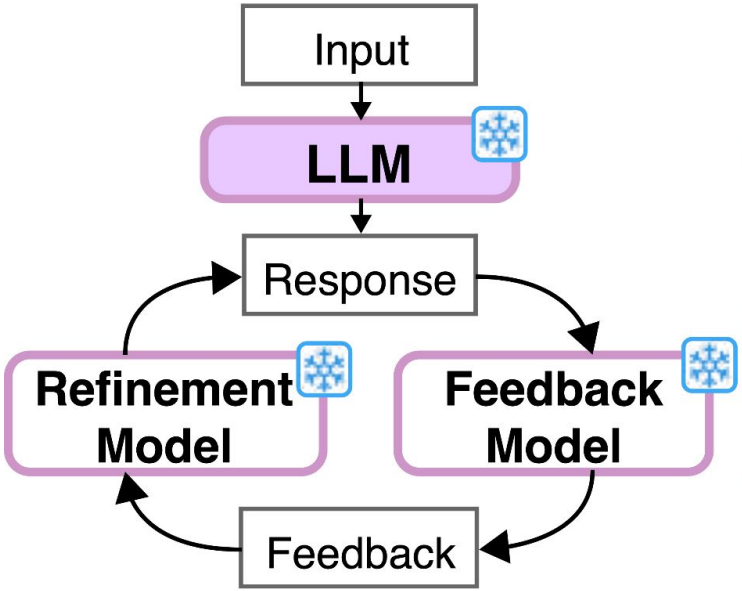
Ryo Kamoi[π], Yusen Zhang[π], Nan Zhang[π], Jiawei Han[ι], Rui Zhang[π]

EMNLP 2024 (TACL)

[π] PennState

[ι] UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

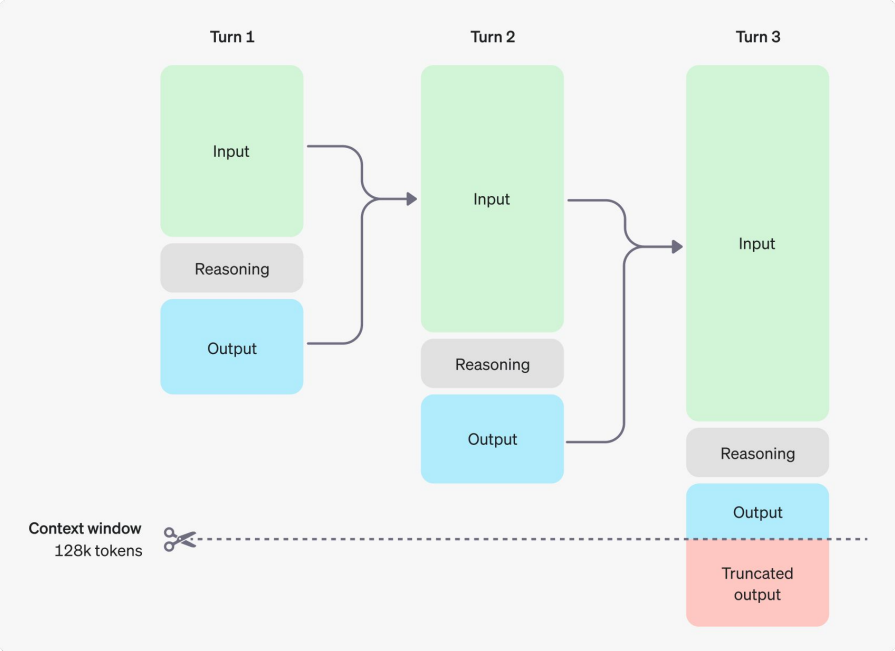
Self-Correction of LLMs



- LLM provides feedback on LLM responses
- LLM refines the responses using the feedback

- Inference-time correction

OpenAI o1



- OpenAI o1 can
  - recognize their mistakes
  - refine their thinking process
  - try different strategies

- One of the first product LLMs with self-correction capabilities

https://openai.com/index/introducing-openai-o1-preview/
https://platform.openai.com/docs/guides/reasoning/how-reasoning-works

Can LLMs Correct Their Own Mistakes?

Our Survey
(originally published in June, 2024)

- Detailed analysis of 40+ papers in self-correction of LLMs
- We conclude that **self-correction is still difficult for LLMs**

Can LLMs Correct Their Own Mistakes?

Our Survey
(originally published in June, 2024)

- Detailed analysis of 40+ papers in self-correction of LLMs
- We conclude that **self-correction is still difficult for LLMs**

OpenAI o1
(released in September 2024)

- **o1 refines its thinking process**
- o1 significantly outperforms GPT-4o on math, coding, science questions, Kaggle, etc.

Can LLMs Correct Their Own Mistakes?

Our Survey
(originally published in June, 2024)

- Detailed analysis of 40+ papers in self-correction of LLMs
- We conclude that **self-correction is still difficult for LLMs**

OpenAI o1
(released in September 2024)

- **o1 refines its thinking process**
- o1 significantly outperforms GPT-4o on math, coding, science questions, Kaggle, etc.

Was our survey wrong...?

Can LLMs Correct Their Own Mistakes?

Our Survey
(originally published in June, 2024)

- Detailed analysis of 40+ papers in self-correction of LLMs
- We conclude that **self-correction is still difficult for LLMs**

OpenAI o1
(released in September 2024)

- **o1 refines its thinking process**
- o1 significantly outperforms GPT-4o on math, coding, science questions, Kaggle, etc.

Was our survey wrong...?

No!

Can LLMs Correct Their Own Mistakes?

Our Survey
(originally published in June, 2024)

- Detailed analysis of 40+ papers in self-correction of LLMs
- We conclude that **self-correction is still difficult for LLMs**

... if we **do not do any training** designed for self-correction

OpenAI o1
(released in September 2024)

- **o1 refines its thinking process**
- o1 significantly outperforms GPT-4o on math, coding, science questions, Kaggle, etc.

... if we do **large-scale reinforcement learning** for self-correction

Takeaways from Our Survey

Be careful about the settings of self-correction!

Be careful about the settings of self-correction!

- Self-correction of LLMs has many hyperparameters in system design
  - Using existing LLMs? **Training** specifically designed for self-correction?

Be careful about the settings of self-correction!

- Self-correction of LLMs has many hyperparameters in system design
  - Using existing LLMs? **Training** specifically designed for self-correction?
  - Using **external tools or information**?

Be careful about the settings of self-correction!

- Self-correction of LLMs has many hyperparameters in system design
  - Using existing LLMs? **Training** specifically designed for self-correction?
  - Using **external tools or information**?
- Some specific **tasks** have favorable properties for self-correction

Takeaways from Our Survey

Be careful about the settings of self-correction!

- Self-correction of LLMs has many hyperparameters in system design
  - Using existing LLMs? **Training** specifically designed for self-correction?
  - Using **external tools or information**?
- Some specific **tasks** have favorable properties for self-correction

Previous papers confuse different settings,
and they sometimes use different settings without explicitly differentiating them!

RQ	Self-Refine (2023)	Huang et al. (2024a)	RCI (2023, §3.1)	RCI (2023, §3.2)	CRITIC (2024, §4.2)	CRITIC (2024, §4.3)	RARR (2023)
RQ1	✓	✗ (§3,5)	✓	–	✗	✗	–
RQ2	–	–	–	✓	✓	✓	–
RQ3	–	✗ (§4)	–	✓	–	✓	✓

Overview

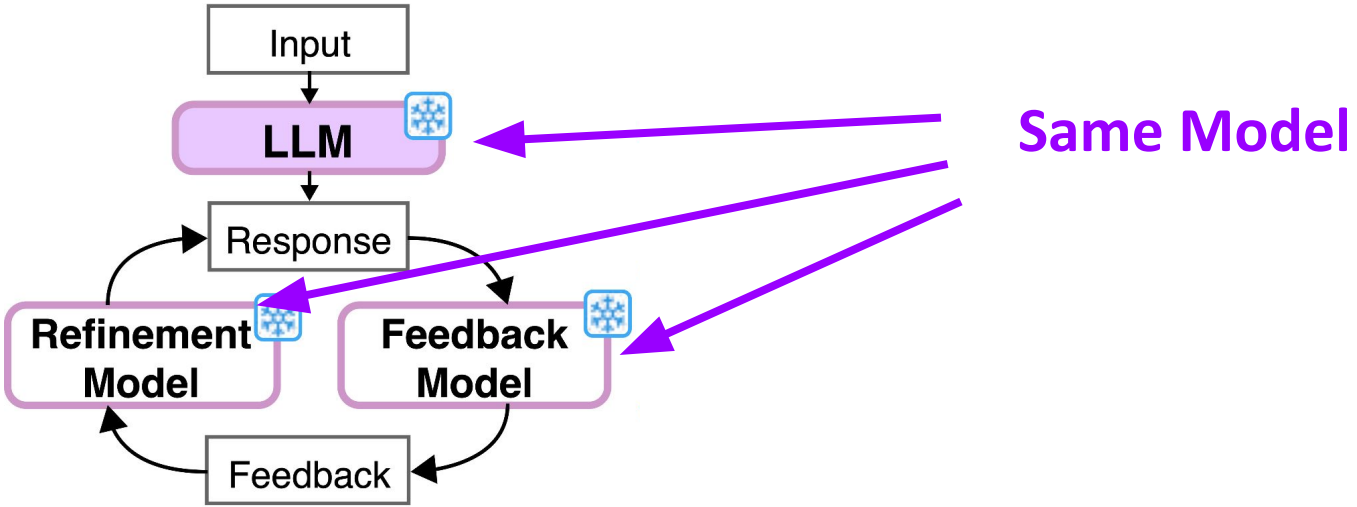
- Intrinsic Self-Correction
- Self-Correction with External Tools and/or Information
- Self-Correction with Additional Training

Overview

- **Intrinsic Self-Correction**
- Self-Correction with External Tools and/or Information
- Self-Correction with Additional Training

Intrinsic Self-Correction

- **Using the same LLMs for initial responses and self-correction**
  - No external tools or information
- Do not train LLMs specifically for self-correction



When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs

Intrinsic Self-Correction: Can LLMs Correct Their Own Mistakes?

- **Intrinsic Self-Correction often does not work** and can degrade performance

Gou et al. (ICLR 2024) “CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing”
Huang et al. (ICLR 2024) "Large Language Models Cannot Self-Correct Reasoning Yet"

Intrinsic Self-Correction: Can LLMs Correct Their Own Mistakes?

- **Intrinsic Self-Correction often does not work** and can degrade performance

Gou et al. (ICLR 2024) “CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing”
Huang et al. (ICLR 2024) "Large Language Models Cannot Self-Correct Reasoning Yet"

- LLMs often cannot detect their own mistakes

Tyen et al. (ACL 2024 Findings) "LLMs cannot find reasoning errors, but can correct them given the error location"
Kamoi et al. (COLM 2024) ”Evaluating LLMs at Detecting Errors in LLM Responses”

Why There are Papers in Successful Intrinsic Self-Correction

However, there are many papers proposing intrinsic self-correction frameworks

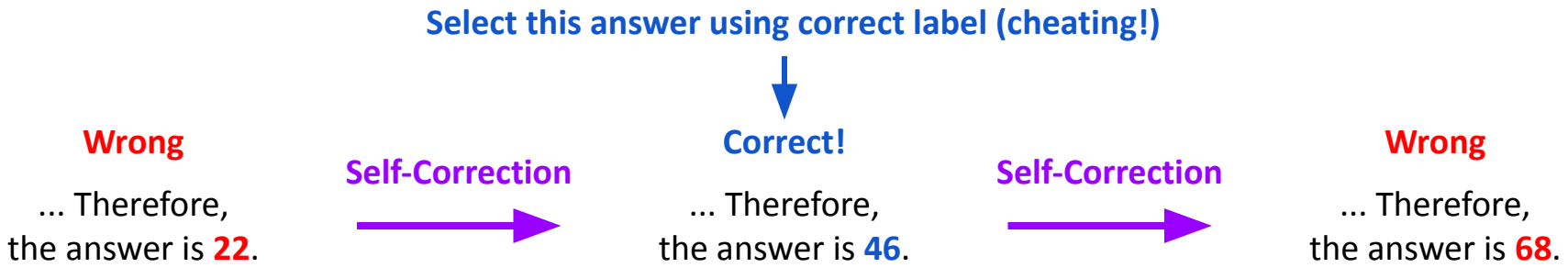
How did they enable intrinsic self-correction?

Why There are Papers in Successful Intrinsic Self-Correction

However, there are many papers proposing intrinsic self-correction frameworks

How did they enable intrinsic self-correction?

- **Using the oracle labels** (i.e., correct answer) – **This is cheating!**
  - Some papers stop self-correction if the current answer is correct
  - This setting ignores the fact that self-correction can degrade responses



Why There are Papers in Successful Intrinsic Self-Correction

However, there are many papers proposing intrinsic self-correction frameworks

How did they enable intrinsic self-correction?

- **Using the oracle labels** (i.e., correct answer) – **This is cheating!**
  - Some papers stop self-correction if the current answer is correct
  - This setting ignores the fact that self-correction can degrade responses

- **Targeting tasks that are exceptionally suitable for self-correction**

Tasks Exceptionally Suitable for Self-Correction

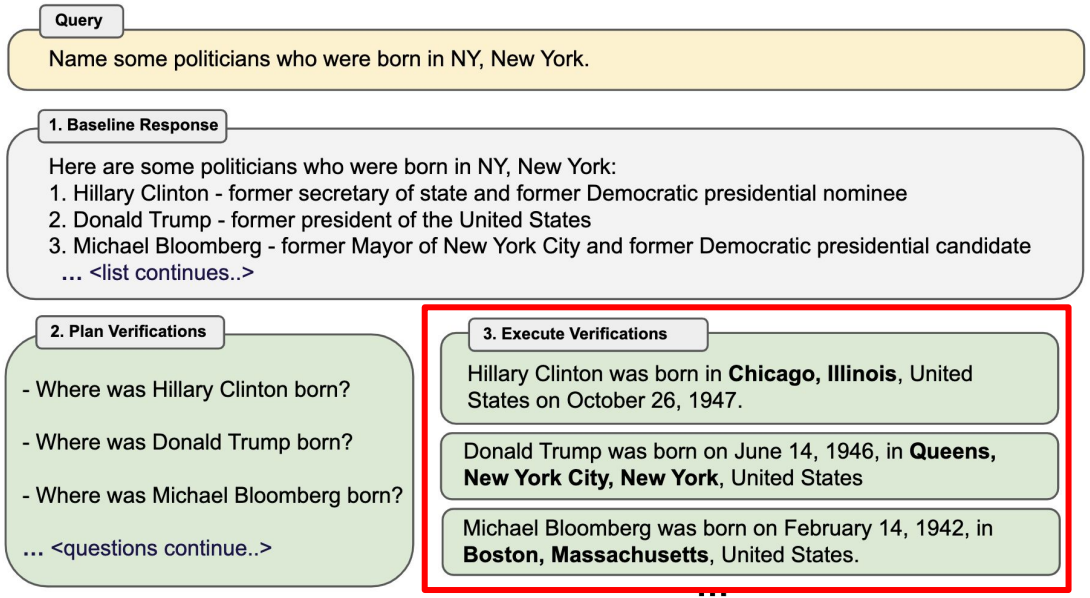
Error detection is a bottleneck of self-correction,
so **intrinsic self-correction is often possible if error detection is easy**

Tasks Exceptionally Suitable for Self-Correction

Error detection is a bottleneck of self-correction,
so **intrinsic self-correction is often possible if error detection is easy**

Decomposable Tasks

- Responses can be decomposed into independent subparts
- Detecting mistakes in each subpart is often easier



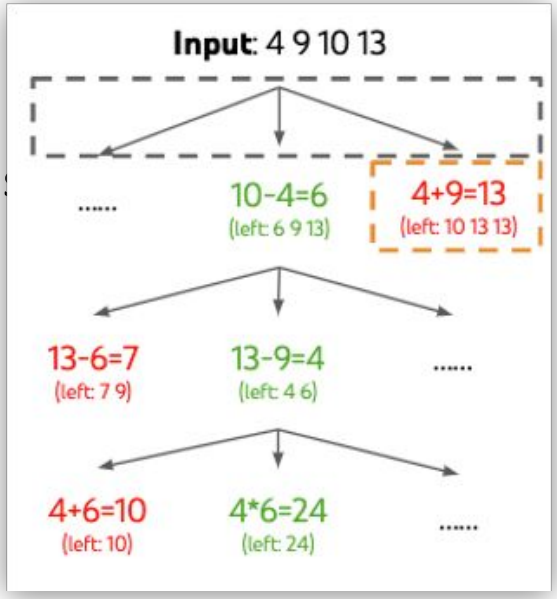
Dhuliawala et al. (ACL 2024 Findings) "Chain-of-Verification Reduces Hallucination in Large Language Models"

Tasks Exceptionally Suitable for Self-Correction

Error detection is a bottleneck of self-correction,
so **intrinsic self-correction is often possible if error detection is easy**

Decomposable Tasks

- Responses can be decomposed into independent
- Detecting mistakes in each subpart is often easier



Verifiable Tasks

- Responses can be evaluated by simple rules
  (e.g., Game of 24)

Yao et al. (NeurIPS 2024) "Tree of Thoughts: Deliberate Problem Solving with Large Language Models"

Intrinsic Self-Correction: Summary

- **Intrinsic Self-Correction is often difficult**
  - LLMs often cannot detect their own mistakes

Intrinsic Self-Correction: Summary

- **Intrinsic Self-Correction is often difficult**
  - LLMs often cannot detect their own mistakes

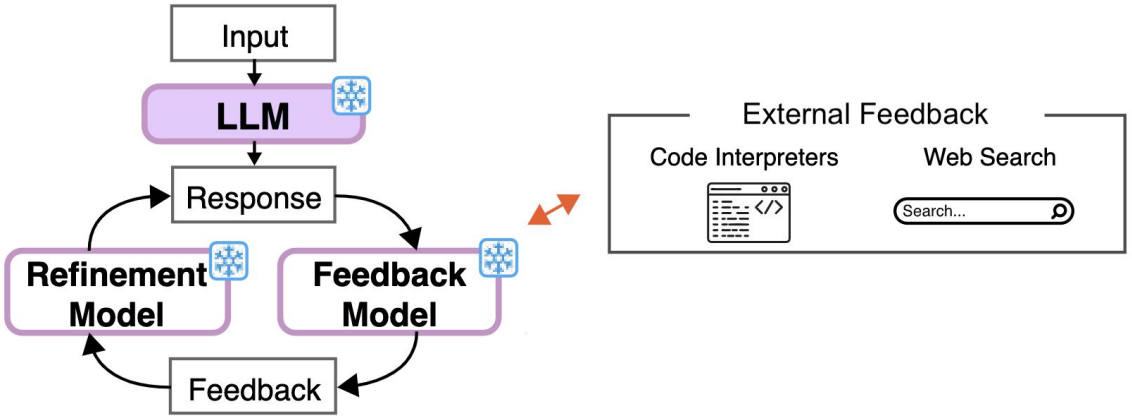
- Some studies use unrealistic settings when evaluating Intrinsic SC
- Some tasks are exceptionally suitable for Intrinsic Self-Correction
  - If error detection is easy, self-correction is often possible
  - But many real-world tasks do not have these properties

Overview

- Intrinsic Self-Correction
- **Self-Correction with External Tools and/or Information**
- Self-Correction with Additional Training

Self-Correction with External Tools or External Information

- We observe that Intrinsic Self-Correction is difficult because LLMs often cannot detect their own mistakes
- Can we improve self-correction if we use **external tools or information**?



When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs

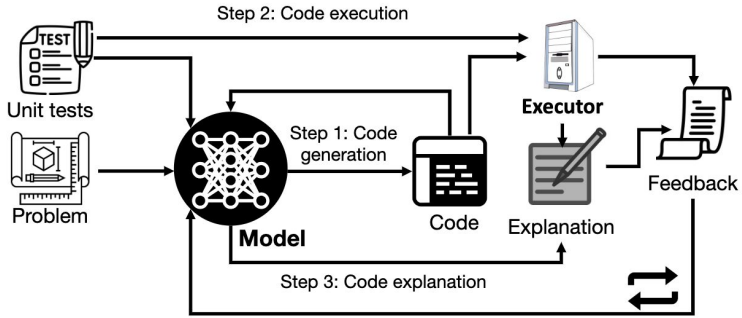
Self-Correction with External Tools

When **tools that can detect mistakes in responses** are available,
LLMs often can refine their own mistakes!

Self-Correction with External Tools

When **tools that can detect mistakes in responses** are available, LLMs often can refine their own mistakes!

Unit tests



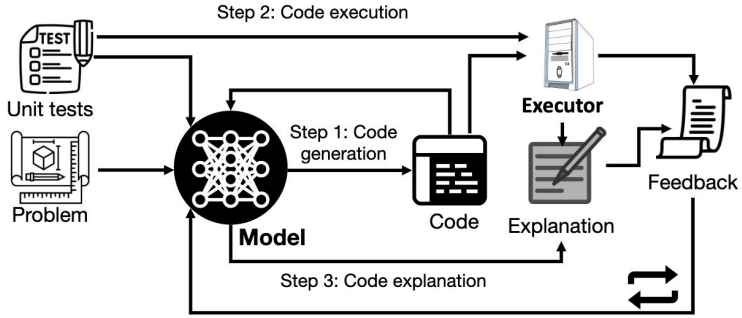
Executor

Chen et al. (ICLR 2024) "Teaching Large Language Models to Self-Debug"

Self-Correction with External Tools

When **tools that can detect mistakes in responses** are available, LLMs often can refine their own mistakes!

Unit tests



Executor

Chen et al. (ICLR 2024) "Teaching Large Language Models to Self-Debug"

Tasks	Tools
Code Generation	Compilers, Code Interpreters
Proof Generation	Proof Assistant
Logical Reasoning	Symbolic Solvers
Simulation Environment	Responses from Environment

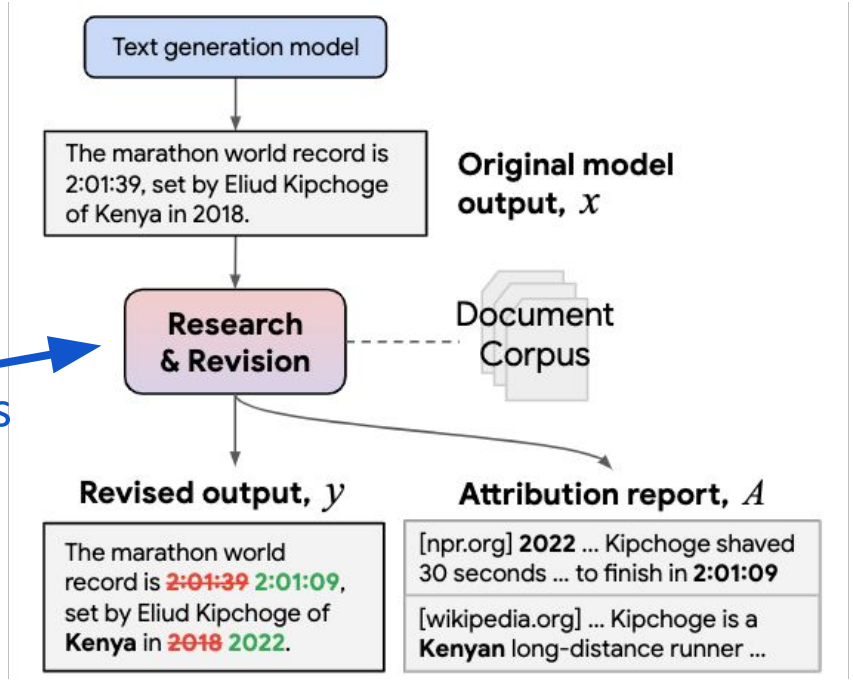
Self-Correction with External Information

When LLMs use **additional information**
during self-correction,
LLMs often can detect and refine
their own mistakes!

Self-Correction with External Information

When LLMs use **additional information** during self-correction, LLMs often can detect and refine their own mistakes!

e.g., Generate queries from initial responses



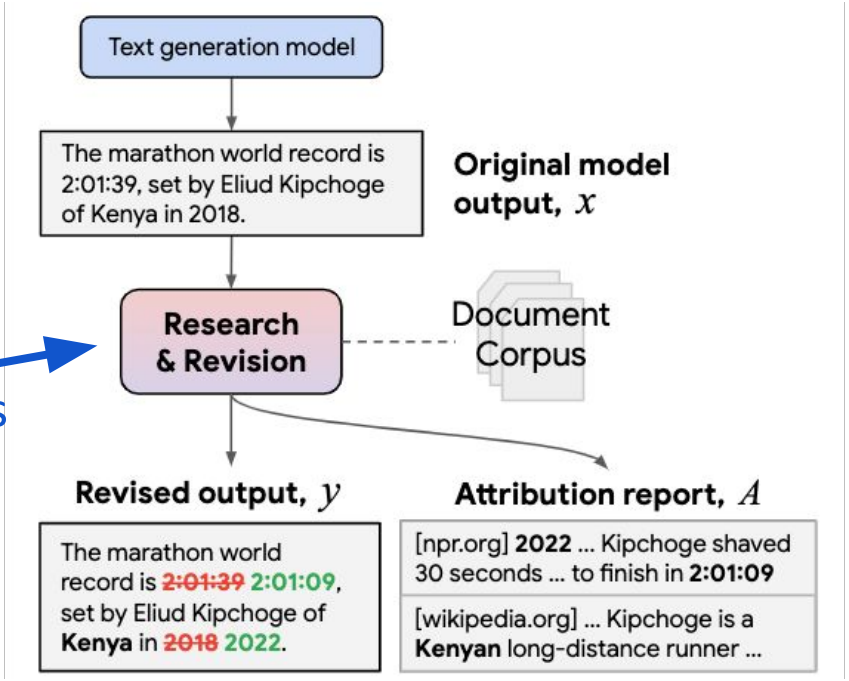
Gao et al. (ACL 2023) "RARR: Researching and Revising What Language Models Say, Using Language Models"

Self-Correction with External Information

When LLMs use **additional information** during self-correction, LLMs often can detect and refine their own mistakes!

e.g., Generate queries from initial responses

- Web search
- Database (e.g., Wikipedia)



Gao et al. (ACL 2023) "RARR: Researching and Revising What Language Models Say, Using Language Models"

Overview

- Intrinsic Self-Correction
- Self-Correction with External Tools and/or Information
- **Self-Correction with Additional Training**

Self-Correction with Training

We concluded that Intrinsic Self-Correction does not work in general tasks

Can we train LLMs for self-correction?

Self-Correction with Training

We concluded that Intrinsic Self-Correction does not work in general tasks

Can we train LLMs for self-correction?

- **Self-correction often works when good training data is available**

Self-Correction with Training

We concluded that Intrinsic Self-Correction does not work in general tasks

Can we train LLMs for self-correction?

- **Self-correction often works when good training data is available**
- There are many methods trains LLMs with SFT or RL for self-correction

Paper	Main Task	Cross-Model	SFT Tasks	Initial Responses		Feedback Generation			Refinement	
				Model	SFT Target	Model	SFT Target	Size	Model	SFT Target
SelFee (2023)	MT-Bench	–	General Tasks	Llama (7B,13B)	ChatGPT Responses	Llama (7B,13B)	ChatGPT Feedback	178K	Llama (7B,13B)	ChatGPT Refinement
Volcano (2024)	Visual Reasoning	–	General Tasks	LLaVA (7B, 13B)	GPT-3.5-T, Human	LLaVA (7B, 13B)	GPT-3.5-T Feedback	274K	LLaVA (7B, 13B)	Reference Answers
Self-Critique (2022)	Topic-based Summarization	–	Target Task	Instruct GPT	Human Summaries	Instruct GPT	Human Feedback	100K	Instruct GPT	Human Refinement
REFINER (2024)	Math, Logic, Moral Stories	✓	Target Task	GPT-3.5	–	T5-base	Synthetic Data	20K - 30K	GPT-3.5	–
Self-Edit (2023b)	Code Generation	✓	Target Task	GPT-3	–	(Code Executor and Test Cases)			PyCodeGPT 110M	Reference Code

Summary

- Intrinsic self-correction often does not work
  - **LLMs often cannot detect and correct their own mistakes**

Summary

- Intrinsic self-correction often does not work
  - **LLMs often cannot detect and correct their own mistakes**
- Specific tasks are suitable for self-correction (e.g., decomposable)

Summary

- Intrinsic self-correction often does not work
  - **LLMs often cannot detect and correct their own mistakes**
- Specific tasks are suitable for self-correction (e.g., decomposable)
- Self-correction is possible when there are good external tools (e.g., coding) or information (e.g., closed book QA)

Summary

- Intrinsic self-correction often does not work
  - **LLMs often cannot detect and correct their own mistakes**
- Specific tasks are suitable for self-correction (e.g., decomposable)
- Self-correction is possible when there are good external tools (e.g., coding) or information (e.g., closed book QA)
- Additional training designed for self-correction is a promising approach
  - OpenAI o1 is (probably) one of the successful models in this approach

Summary

- Intrinsic self-correction often does not work
  - **LLMs often cannot detect and correct their own mistakes**
- Specific tasks are suitable for self-correction (e.g., decomposable)
- Self-correction is possible when there are good external tools (e.g., coding) or information (e.g., closed book QA)
- Additional training designed for self-correction is a promising approach
  - OpenAI o1 is (probably) one of the successful models in this approach

When comparing performance of self-correction frameworks, be careful about the settings!

Summary

- Intrinsic self-correction often does not work
  - **LLMs often cannot detect and correct their own mistakes**
- Specific tasks are suitable for self-correction (e.g., decomposable)
- Self-correction is possible when there are good external tools (e.g., coding) or information (e.g., closed book QA)
- Additional training designed for self-correction is a promising approach
  - OpenAI o1 is (probably) one of the successful models in this approach

When comparing performance of self-correction frameworks, be careful about the settings!

Please reach out us for any questions! **Ryo Kamoi: ryokamoi@psu.edu**