When Hindsight is Not 20/20: Testing Limits on Reflective Thinking in Large Language Models

Anonymous ACL submission

Abstract

Recent studies suggest that self-reflective prompting can significantly enhance the reasoning capabilities of Large Language Models (LLMs). However, the use of external feedback as a stop criterion raises doubts about the true extent of LLMs' ability to emulate human-like self-reflection. In this paper, we set out to clarify these capabilities under a more stringent evaluation setting in which we disallow any kind of external feedback. Our findings under this setting show a split: while self-reflection enhances performance in TruthfulQA, it adversely affects results in HotpotQA. We conduct follow-up analyses to clarify the contributing factors in these patterns, and find that the influence of self-reflection is impacted both by reliability of accuracy in models' initial responses, and by overall question difficulty: specifically, self-reflection shows the most benefit when models are less likely to be correct initially, and when overall question difficulty is higher. We also find that self-reflection reduces tendency toward majority voting. Based on our findings, we propose guidelines for decisions on when to implement self-reflection.

1 Introduction

Large Language Models (LLMs) have shown impressive performance in generating human-like text (e.g., ChatGPT (OpenAI, 2021)), and recent works demonstrate that we can further prompt LLMs to reflect on their own outputs to improve their capabilities on complicated reasoning, programming and planning tasks (Huang et al., 2022; Kim et al., 2023; Madaan et al., 2023; Shinn et al., 2023; Chen et al., 2023b; Wang et al., 2023b) and also improve their alignment with human values (e.g., less harmful and more helpful) (Bai et al., 2022; Ganguli et al., 2023).[1] However, Huang et al. (2023) find that

[1] Various terms like "self-reflection", "self-refine", "self-correction", and "self-improvement" describe these introspective behaviors. For clarity and consistency, we will exclusively use "self-reflection" in this paper.

Step 1: Exploration **Repeat K times**
Instructions: [Task-Specific Instruction]
Question: [Insert Question]
Answer:
Step 2: Reflection **Repeat K times**
Instructions: [Task-Specific Instruction]
Please critique your answer based on the given question.
Question: [Insert Question]
Answer: [Insert Answer from Exploration]
Reflection:
Step 3: Revision
Instructions: [Task-Specific Instruction]
Question: [Insert Question]
Answer: [Insert Answer from Exploration]
Reflection: [Insert Reflection]
{Concatenate all K responses and reflection}
Question: [Insert Question]
Answer:

Figure 1: Example of Self-Reflection Prompting

performance gains associated with self-reflection may be due to implicit usage of external feedback as a stop criterion, as well as overly-engineered prompts that bias the model outputs, casting doubt on the true effectiveness of self-reflection.

To verify the extent to which LLMs can truly reflect on their outputs, we take a more stringent evaluation approach: in addition to excluding external feedback (Huang et al., 2023), we also disallow multi-round iterative prompting, which can hint to the model that its prior response is incorrect. Instead, we sample multiple model responses given a prompt, and ask the model to self-reflect on these candidate outputs. With this *single-round testing*, we can zero in on the model's ability to use self-reflection without implicit hints about whether a given response candidate is correct or incorrect.

Our experiments show that, in a case study with ChatGPT on different QA datasets, self-reflection in our setting yields mixed results. Specifically, self-reflection improves performance on TruthfulQA (Lin et al., 2022), but decreases model performance in HotpotQA (Yang et al., 2018). Through follow-up analyses, we identify that the effectiveness of self-reflection strongly depends on the confidence in accuracy of the model's initial responses, as well as overall question difficulty as judged by humans: when the model is reliably giving correct answers from the start, self-reflection is more often harmful—however, on questions of greater difficulty, self-reflection is beneficial even when a decent percent of initial model responses are correct. We also find that self-reflection reduces model tendency toward majority voting, suggesting more sophisticated decision-making (albeit sometimes resulting in lower accuracy). Based on our findings, we propose a practical guideline for users to decide when to use self-reflection.

2 Self-Reflection Prompting

To focus on evaluating intrinsic reflective thinking capability, we adopt the following evaluation setting: in addition to the Huang et al. (2023) protocol of excluding external feedback and prompt optimization, we additionally disallow *iterative prompting*, which samples new responses based on previous responses, creating an implicit hint to bias the model behavior (Huang et al., 2023).[2] We call our approach *Single-Round Self-Reflection Verification (SR^{2}V)*. We evaluate LLMs' reflective thinking capability using the following simple three-stage format: 1) *Exploration:* Given an input X, we prompt LLM M to generate K candidate responses $r_j \sim P_M(r_j|X, I_{\text{Exploration}}), 1 \leq j \leq K$ with instruction $I_{\text{Exploration}}$. 2) *Reflection:* For each response r_j, we prompt M with the concatenated input $[X; r_j]$ to generate a self-critique $c_j \sim P_M\left(c_j|[X; r_j], I_{\text{Reflection}}\right)$ with another instruction $I_{\text{Reflection}}$. 3) *Revision:* We concatenate the K response-reflection pairs into a new input and prompt M to generate an improved output. An illustration of this procedure is shown in Figure 1.

3 Preliminary Study: Does Self-Reflection Prompting Work Under SR^{2}V?

We follow previous works (Bai et al., 2022; Shinn et al., 2023; Huang et al., 2023) in using two representative datasets, TruthfulQA and HotpotQA, to verify the effectiveness of self-reflection under SR^{2}V. TruthfulQA is designed to evaluate the truthfulness of LMs' responses, while HotpotQA focuses on multi-hop reasoning tasks, aimed at requiring complex reasoning capabilities.

Experiment Setup For these experiments we set $K = 4$, and we prompt ChatGPT-3.5 ("gpt-3.5-turbo-16k-0613") with the questions from each dataset.[3] For TruthfulQA we evaluate automatically (see details in Appendix D). For HotpotQA, we find that traditional exact match often unfairly

[2]We present a performance comparison between iterative prompting and non-iterative prompting in Appendix C.

[3]The 16k variant is chosen to accommodate responses and reflection pairs that exceed the standard 4096 token limit, particularly in detailed experiments of Section 5.

Metric	Standard Prompting	Exploration-Only	Self-Reflection
	TruthfulQA		
Rouge-1	57.5 ± 1.1	57.2	**60.8**
BLEURT	66.8 ± 1.9	60.7	**72.8**
	HotpotQA		
Accuracy*	80.3 ± 0.5	**80.8**	76.2
EM	$\mathbf{50.5 \pm 0.4}$	47.3	37.0

Table 1: Self-reflection SR^{2}V experiment results on QA datasets. Bold-facing indicates the best-performing method under each metric. *Evaluated manually.

assigns 0 score for semantically correct model responses; therefore, we manually assess $1,000$ randomly chosen HotpotQA instances to check the model's answers against references. All prompt templates used can be found in Appendix E. To isolate the specific effect of the generated reflections, we also include an **exploration-only** baseline, in which we keep the Exploration but remove the Reflection component, and only concatenate the candidate model responses in the Revision prompt.[4]

Observations The results are shown in Table 1. In TruthfulQA, we see that using self-reflection achieves significantly better performance than either the exploration-only baseline or standard prompting. This finding is consistent with the observation of Bai et al. (2022) that LLMs' self-evaluation (in the form of reflection) can help to produce more factual outputs. However, we see that on HotpotQA, accuracy when using self-reflection is about 4% worse compared to both the exploration-only baseline and standard prompting. These results suggest that self-reflection may in fact harm performance in multi-hop reasoning tasks. This aligns with the self-reflection limitations found in Huang et al. (2023), and verifies that these limitations also extend to our more stringent evaluation setting, but presents a more complicated picture with the continued effectiveness of self-reflection on TruthfulQA under this setting.

4 Why Self-Reflection May Not Work?

To better understand these patterns, we conduct an error analysis drawing inspiration from the re-

[4]The exploration-only baseline can be viewed as one implementation of (universal) self-consistency prompting (Wang et al., 2023a; Chen et al., 2023a). Rather than applying majority voting directly to the outputs, this method involves inputting these outputs back into the model for aggregation. As we'll explore in Section 6, we also find the model predominantly engages in a form of majority voting in this process.

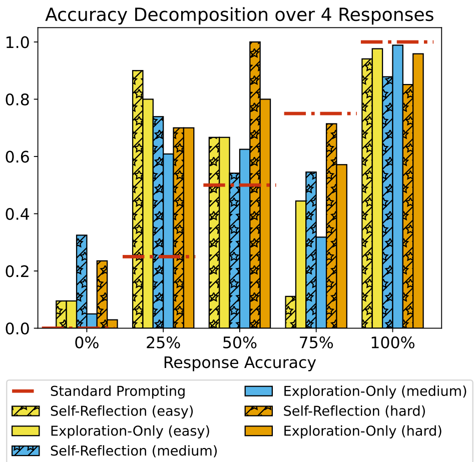

Figure 2: Performance Decomposition on Question Difficulty and Response Accuracy.

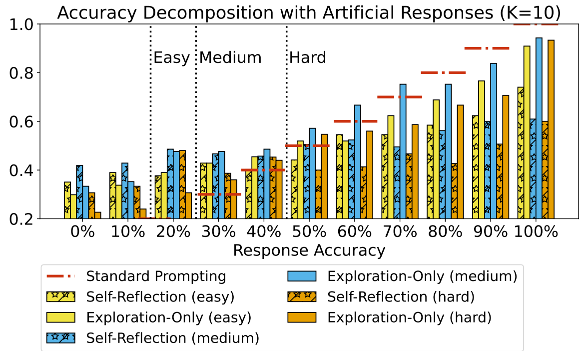

Figure 3: Performance Decomposition on Question Difficulty and Response Accuracy (Artificial Responses). Dotted lines show "turning points" at which reflection loses effectiveness, for Easy/Medium/Hard questions.

flection conceptual model in psychology (Hommel et al., 2023). We hypothesize that two key factors influence self-reflection's efficacy: 1) the objective **question difficulty** (quantifiable based on human annotations), and 2) the **model's comprehension quality** (quantifiable based on the proportion of correct responses). Following this framework, we can predict that if a question is above average in human-annotated difficulty, self-reflection may be of greater benefit. Similarly, if the model already has a strong grasp of the question, it may not benefit as much from self-reflection.

To test these hypotheses, we break down model performance based on levels of question difficulty and model comprehension. We focus on HotpotQA, as human judgments of question difficulty are available as annotations in this dataset, and this dataset also enables a clearly-defined notion of accuracy. We use these human difficulty annotations for question difficulty, and for model comprehension we use Response Accuracy (RA): the proportion of correct answers among the K candidate model responses sampled during Exploration.

The broken-down results are shown in Figure 2. The results show an interaction between our two variables. For questions judged by humans as Easy, self-reflection shows a benefit only when the model's candidate responses are mostly—but not all—incorrect, with self-reflection otherwise having negligible or negative effects on performance. For questions judged as Medium, there is a more even split: when most or all of the model's candidate responses are wrong, self-reflection is beneficial, but when half or more of the responses are correct, self-reflection is often harmful—with the notable exception of the 75% RA bin. A similar pattern is seen for questions judged as Hard, though for this category self-reflection is more consistently beneficial through the 75% RA bin, showing harm to performance only when all candidate model responses are already correct.

5 Error Analysis via Artificial Response

The above analysis suggests an interaction between difficulty and comprehension variables in effectiveness of self-reflection—however, our ability to disentangle these effects is limited by imbalanced distribution of model comprehension relative to question difficulty. To assess the interaction more thoroughly, we simulate model "mis-comprehension" across a wider range of question difficulties, by sampling model responses to minimally edited versions of the prompts, and then pairing these responses with the original prompts when eliciting self-reflection. This allows us to increase the number of incorrect candidate responses, and thus to more evenly distribute RA levels across human difficulty levels. More details on this simulation process can be found in Appendix B.

For this experiment, we generate $K = 10$ candidate responses per question, with a mix of synthetic pairings and real pairings.[5] Results are shown in Figure 3. We see that the benefits of self-reflection are now limited to the lowest RA levels, and there is also now a clearer shift from beneficial to harmful effects of self-reflection as RA increases. We also see that the interaction with question difficulty remains: the turning point from beneficial to harmful falls around 50% RA for Hard questions, 30% for Medium questions, and 20% for Easy questions. Overall, this indicates that a major contributor to

[5]We also plot the performance decomposition over K=4 artificial responses in Appendix A.

the effectiveness of self-reflection is the confidence of model accuracy on the question—if the model is reliably correct on initial responses, self-reflection tends to be harmful. However, this effect is further modulated by overall question difficulty: the benefits of self-reflection persist to higher levels of response accuracy if the questions are more difficult based on human judgment.

Though TruthfulQA is not as conducive to exact quantification of our variables, based on these results we can now speculate that the effectiveness of self-reflection on that dataset may be attributable to lower rate of good initial model responses, and potentially also higher overall question difficulty.

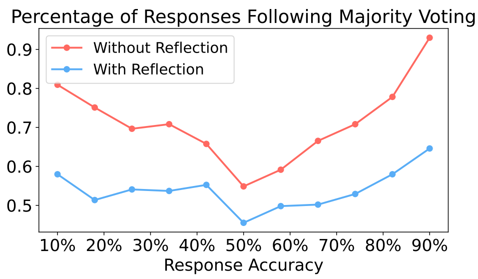

Figure 4: Majority Voting Analysis

6 Effects on majority voting

A natural question to ask at this point is to what extent the effect of RA is due to the model employing majority voting on the candidate responses. In Figure 4 we plot the percentage of items in which the model's output is consistent with majority voting, at different RA levels (computed at $K = 10$ including artificially generated responses), both with and without self-reflection. The plot shows that without self-reflection, the tendency to give answers consistent with majority voting is strong and closely correlated with the strength of the accuracy trend (i.e., more majority voting when most candidate responses are either correct or incorrect, and less majority voting when candidates are more mixed). However, *with* self-reflection the tendency to align with majority voting is significantly reduced across RA levels, suggesting that self-reflection does encourage more sophisticated decision strategies (even if in the case of higher RA levels, this in fact has a harmful effect on accuracy).

7 Discussion

Our analyses above have found that self-reflection benefits are limited to cases in which model accuracy is unreliable on initial responses, though benefits are more persistent for harder questions. Based on these findings, we propose a set of guidelines for determining when to implement self-reflection in practical applications, for a given request or prompt. The core principle involves basing decisions on estimated RA and question difficulty, and these guidelines can be applied by simply sampling responses for the target question or prompt. First, if external tools or certain access to ground truth answers are available such that RA can be reliably estimated, then self-reflection should be used when RA levels are low. Next, if difficulty annotations/subjective difficulty judgements are available, self-reflection can also be promising when RA levels are intermediate and question difficulty is high. If RA cannot be estimated, response consistency can be used as a proxy: if responses are highly consistent, self-reflection may be unlikely to provide benefit. If consistency is low, then self-reflection may be beneficial, especially for questions of higher difficulty. An illustration of these guidelines is in Figure 5.

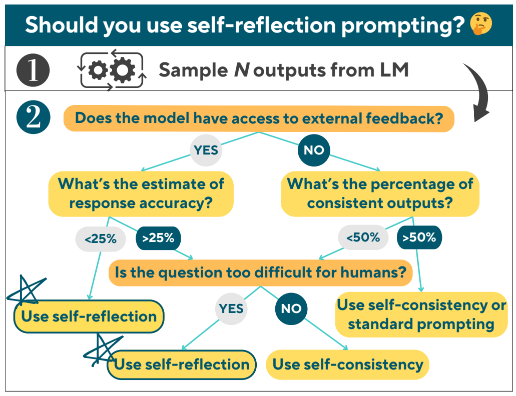

Figure 5: Proposed guide for using Self-Reflection.

8 Conclusion

In this paper, we evaluate ChatGPT's self-reflective capabilities under a stringent single-round multi-response evaluation setting. We find mixed results, and further analysis shows that the effectiveness of self-reflection is impacted both by question difficulty and by model response accuracy level: benefits of self-reflection are mostly limited to cases in which the model's initial responses are unreliable in accuracy, but with more persistent benefits for harder questions. Additionally, we find that self-reflection reduces the model's tendency for majority voting. We propose guidelines for when to use self-reflection, and we look forward to work further exploring impacts on self-reflection, and further refining these guidelines.

Limitations

In this work, we adopt a stringent evaluation strategy to test the effectiveness of self-reflective abilities of LLMs. One limitation is that our experiments are all based on a snapshot of the ChatGPT model (gpt-3.5-turbo-16k-0613). We focus on ChatGPT because it is a state-of-the-art (SOTA) chat model, and it allows us to make our results directly comparable with previous work. We only examine one model to ensure that results will not be affected by model updates. However, the assessment of self-reflection may vary between different versions of ChatGPT, as well as between ChatGPT and other LLMs.

Secondly, we use only two datasets for evaluating reflective ability. We chose these two datasets for a focused study covering two very different QA domains, but we look forward to future work further extending these types of analyses to a broader collection of datasets.

Thirdly, we conducted an artificial response experiment in Section 5 to simulate the real output distribution of the language model. This is a rough estimate of ChatGPT's actual output distribution. As we sampled ten fake responses from the language model, it is impossible to cover all possible cases of outputs, and there might be bias in the sample distribution. Future work could try generating a higher number of fake responses to obtain a more accurate distribution of the model.

Finally, although RA proves a valuable metric for determining the utility of self-reflection, its reliance on access to ground truth undermines its practical use. An initial attempt to use GPT-4 to produce an estimate of RA yielded unsatisfactory results (detailed in Appendix F). Further examination of this topic is reserved for future research.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023a. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023b. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Mandy Hommel, Bärbel Fürstenau, and Regina H Mulder. 2023. Reflection at work–a conceptual model and the meaning of its components in the domain of VET teachers. *Frontiers in Psychology*, 13:923888.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

OpenAI. 2021. Chatgpt. `https://openai.com/api/models/gpt-chat/`.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models.

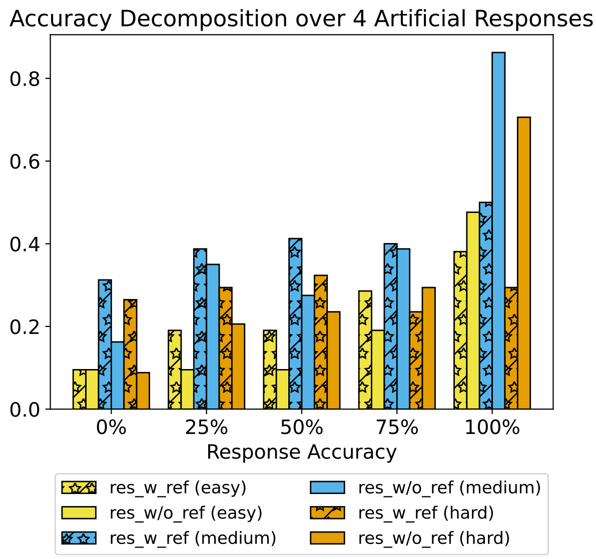

Figure 6: Accuracy vs. Correctness Margin for each artificial response

Ziqi Wang, Le Hou, Tianjian Lu, Yuexin Wu, Yunxuan Li, Hongkun Yu, and Heng Ji. 2023b. Enable language models to implicitly learn self-improvement from data. *arXiv preprint arXiv:2310.00898*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

A Accuracy Decomposition over 4 responses

See Figure 6.

B Artificial Response Generation

We do artificial response generation by prompting ChatGPT to edit the context used in HotpotQA. Specifically, the following steps were adopted: 1) For chosen questions, perform a simple perturbation on the context (e.g., entity replacement). An example is shown in Figure 7. 2) Manually inspect some samples to ensure minimal edits and answerability. 3) Prompt the model to regenerate responses and reflections based on the altered context. In this way, we are simulating scenarios where the model doesn't comprehend the context perfectly. [6]

[6]While directly editing outputs to create correct or incorrect answers is an option, we avoid this to ensure the results reflect the model's natural response distribution.

Here is an example for how we modify the context:

Original question: What nationality was James Henry Miller's wife?

Original context: ... Ewan MacColl: James Henry Miller (25 January 1915 – 22 October 1989), better known by his stage name Ewan MacColl, was an English folk singer, songwriter, communist, labour activist, actor, poet, playwright and record producer. Peggy Seeger: Margaret "Peggy" Seeger (born June 17, 1935) is an American folksinger. She is also well known in Britain, where she has lived for more than 30 years, and was married to the singer and songwriter Ewan MacColl until his death in 1989. ...

Fake context 1: ... Ewan MacColl: James Henry Miller (25 January 1915 – 22 October 1989), better known by his stage name Ewan MacColl, was a Scottish folk singer, songwriter, capitalist, labour activist, actor, poet, playwright and record producer.. Peggy Seeger: Margaret "Peggy" Seeger (born June 17, 1935) is an American country singer. She is also well known in France, where she has lived for more than 30 years, and was married to the actor and playwright Ewan MacColl until his death in 1989. ...

Fake context 2: ... Ewan MacColl: James Henry Miller (25 January 1915 – 22 October 1989), better known by his stage name Ewan MacColl, was an Australian folk singer, songwriter, conservative, labour activist, actor, poet, playwright and record producer. Peggy Seeger: Margaret "Peggy" Seeger (born June 17, 1935) is a British pop singer. She is also well known in Germany, where she has lived for more than 30 years, and was married to the musician and producer Ewan MacColl until his death in 1989. ...

Fake context 3: ... Ewan MacColl: James Henry Miller (25 January 1915 – 22 October 1989), better known by his stage name Ewan MacColl, was a Canadian folk singer, songwriter, anarchist, labour activist, actor, poet, playwright and record producer. Peggy Seeger: Margaret "Peggy" Seeger (born June 17, 1935) is an American rapper. She is also well known in Spain, where she has lived for more than 30 years, and was married to the actor and politician Ewan MacColl until his death in 1989. ...

Fake context 4: ... Ewan MacColl: James Henry Miller (25 January 1915 – 22 October 1989), better known by his stage name Ewan MacColl, was an

Irish folk singer, songwriter, monarchist, labour activist, actor, poet, playwright and record producer. Peggy Seeger: Margaret "Peggy" Seeger (born June 17, 1935) is a French jazz singer. She is also well known in Italy, where she has lived for more than 30 years, and was married to the artist and filmmaker Ewan MacColl until his death in 1989. ...

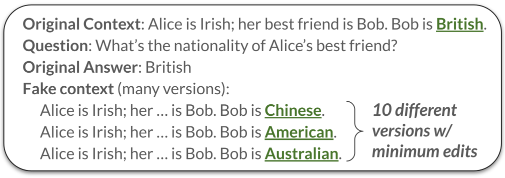

Figure 7: Synthesized Artificial Contexts Example

Metric	Standard Prompting	Exploration-Only	Self-Reflection
	TruthfulQA		
Rouge-1	57.5 ± 1.1	55.1	**59.0**
BLEURT	66.8 ± 1.9	70.1	**72.9**
	HotpotQA		
Accuracy	80.2 ± 0.4	69.7	71.9

Table 2: Self-Reflection experiment results using iterative prompting. Bold-faced numbers at each row indicate the best-performing method under each metric.

C Conditional Prompting Results

We demonstrate the conditional prompting results in Table 2. Comparing the results in Table 1 and Table 2, we can see that there is no significant difference between these parallel prompting and conditional prompting. To avoid the implicit bias introduced by conditional prompting, as Huang et al. (2023) point out, we stick to parallel prompting to conduct our evaluation on self-reflective thinking capability.

D Evaluation details for TruthfulQA

We use the generation setting of TruthfulQA, which evaluates by comparing how closely the model's responses match a preferred reference versus an undesired one We follow (Lin et al., 2022) to use Rouge-1 (Lin, 2004) and BLEURT (Sellam et al., 2020) for similarity computation.

E Prompts used in Experiment

E.1 TruthfulQA: Standard Prompt

```
messages=[
    {"role": "user",
    "content": question}
]
```

E.2 TruthfulQA: Response Critique Prompt

```
messages=[
    {"role": "system",
    "content": "You are a helpful
    assistant."},
    {"role": "user",
    "content": question},
    {"role": "assistant",
    "content": response},
    {"role": "user",
    "content": "Could you critique
    your last response?"}
]
```

E.3 TruthfulQA: Response Without Reflection

```
messages=[
    {"role": "system",
    "content": "You are a helpful
    assistant."},
    {"role": "user",
    "content": question},
    {"role": "assistant",
    "content": response_1},
    {"role": "user",
    "content": question},
    {"role": "assistant",
    "content": response_2},
    {"role": "user",
    "content": question},
    {"role": "assistant",
    "content": response_3},
    {"role": "user",
    "content": question}
]
```

E.4 TruthfulQA: Response With Reflection

```
messages=[
    {"role": "system",
    "content": "You are a helpful
    assistant."},
    {"role": "user",
    "content": question},
    {"role": "assistant",
    "content": response_1},
    {"role": "user",
    "content": "Please critique your
    responses"},
    {"role": "assistant",
    "content": critique_1},
    {"role": "user",
    "content": question},
    {"role": "assistant",
    "content": response_2},
    {"role": "user",
    "content": "Please critique your
    responses"},
    {"role": "assistant",
    "content": critique_2},
    {"role": "user",
    "content": question},
    {"role": "assistant",
    "content": response_3},
    {"role": "user",
    "content": "Please critique your
```

```
        responses"},
        {"role": "assistant",
        "content": critique_3},
        {"role": "user",
        "content": question}
    ]
```

E.5 HotpotQA: Standard Prompt

```
messages=[
    {"role": "system",
    "content": "You are a helpful
    assistant. Answer the question
    based on the context provided.
    Provide extremely concise answers
    with no explanation."},
    {"role": "user",
    "content": "Context: Earth: The
    Earth is the third planet from
    the Sun. Question: Which planet
    is Earth from the Sun? Answer:
    Third"},
    {"role": "user",
    "content": f"Context:
    {formatted_context}\n
    Question: {question}\nProvide a
    short answer without
    explanation."}
]
```

E.6 HotpotQA: Response Critique Prompt

```
messages=[
    {"role": "system",
    "content": "You are a helpful
    assistant. Answer the question
    based on the context provided."},
    {"role": "user",
    "content": f"Context:
    {formatted_context}\n
    Question: {question}"},
    {"role": "assistant",
    "content": f"{response}"},
    {"role": "user",
    "content": f"Please review and
    critique your previous response,
    and keep in mind not to add any
    unnecessary apologies. You can
    refer back to the original
    context if needed."}
]
```

E.7 HotpotQA: Response Without Reflection

```
messages=[
    {"role": "system",
    "content": "You are a helpful
    assistant. Answer the question
    based on the context provided.
    Provide extremely concise answers
    with no explanation."},
    {"role": "user",
    "content": "Context: Earth: The
    Earth is the third planet from
    the Sun. Question: Which planet
    is Earth from the Sun?
    Answer: Third"},
    {"role": "user",
    "content": f"Context:
    {formatted_context}\n
    Question: {question}\n
    Provide a short answer without
    explanation."},
    {"role": "assistant",
    "content": f"{response_1}"},
    {"role": "user",
    "content": f"{question}\n
    Provide a short answer without
    explanation."},
    {"role": "assistant",
    "content": f"{response_2}"},
    {"role": "user",
    "content": f"{question}\n
    Provide a short answer without
    explanation."},
    {"role": "assistant",
    "content": f"{response_3}"},
    {"role": "user",
    "content": f"{question}\n
    Provide a short answer without
    explanation."},
    {"role": "assistant",
    "content": f"{response_4}"},
    {"role": "user",
    "content": f"{question}\n
    Provide a short answer without
    explanation."},
]
```

E.8 HotpotQA: Response With Reflection

```
messages=[
    {"role": "system",
    "content": "You are a helpful
    assistant. Answer the question
    based on the context provided.
    Provide extremely concise answers
    with no explanation."},
    {"role": "user",
    "content": "Context: Earth: The
    Earth is the third planet from the
    Sun. Question: Which planet is Earth
    from the Sun? Answer: Third"},
    {"role": "user",
    "content": f"Context:
    {formatted_context}\n
    Question: {question}\n
    Provide a short answer without
    explanation."},
    {"role": "assistant",
    "content": f"{response_1}"},
    {"role": "user",
    "content": f"Please review and
    critique your previous response,
    and keep in mind not to add any
    unnecessary apologies. You can
    refer back to the original context
    if needed."},
    {"role": "assistant",
    "content": f"{critique_1}"},
    {"role": "user",
    "content": f"{question}\n
    Provide a short answer without
    explanation."},
    {"role": "assistant",
    "content": f"{response_2}"},
    {"role": "user",
```

```
        "content": f"Please review and
        critique your previous response,
        and keep in mind not to add any
        unnecessary apologies. You can
        refer back to the original context
        if needed."},
        {"role": "assistant",
        "content": f"{critique_2}"},
        {"role": "user",
        "content": f"{question}\n
        Provide a short answer without
        explanation."},
        {"role": "assistant",
        "content": f"{response_3}"},
        {"role": "user",
        "content": f"Please review and
        critique your previous response,
        and keep in mind not to add any
        unnecessary apologies. You can
        refer back to the original
        context if needed."},
        {"role": "assistant",
        "content": f"{critique_3}"},
        {"role": "user",
        "content": f"{question}\n
        Provide a short answer without
        explanation."},
        {"role": "assistant",
        "content": f"{response_4}"},
        {"role": "user",
        "content": f"Please review and
        critique your previous response,
        and keep in mind not to add any
        unnecessary apologies. You can
        refer back to the original context
        if needed."},
        {"role": "assistant",
        "content": f"{critique_4}"},
        {"role": "user",
        "content": f"{question}\n
        Provide a short answer without
        explanation."}
    ]
```

E.9 HotpotQA: Fake Evidence Generation

```
messages=[
    {"role": "system",
    "content": "You are a helpful
    assistant."},
    {"role": "user",
    "content": f"Here is a question:
    {question}. Please create 10
    different versions of 'fake
    supporting facts' based on the
    following real supporting facts.
    Modify only one sentence in each
    version, making sure the modified
    sentence is still relevant but
    contains false information. Keep
    the other sentences unmodified.
    Each version of fake supporting
    facts should have the same number
    of sentences as the real
    supporting facts."},
    {"role": "user",
    "content": f"Real Supporting
    Facts:{real_sf}"},
    {"role": "user",
    "content": "Please generate the
    fake supporting facts versions.
    Remember to index all the sentences.
    You must generate 10 versions
    before you stop."},
    {"role": "user",
    "content":
    f"Fake Supporting Facts Version 1:\n
    [Insert manipulated sentences here]\
        ↪ n
    Fake Supporting Facts Version 2:\n
    [Insert manipulated sentences here]\
        ↪ n
    Fake Supporting Facts Version 3:\n
    [Insert manipulated sentences here]\
        ↪ n
    Fake Supporting Facts Version 4:\n
    [Insert manipulated sentences here]\
        ↪ n
    Fake Supporting Facts Version 5:\n
    [Insert manipulated sentences here]\
        ↪ n
    Fake Supporting Facts Version 6:\n
    [Insert manipulated sentences here]\
        ↪ n
    Fake Supporting Facts Version 7:\n
    [Insert manipulated sentences here]\
        ↪ n
    Fake Supporting Facts Version 8:\n
    [Insert manipulated sentences here]\
        ↪ n
    Fake Supporting Facts Version 9:\n
    [Insert manipulated sentences here]\
        ↪ n
    Fake Supporting Facts Version 10:\n
    [Insert manipulated sentences here
        ↪ ]"},
]
```

F Challenges in Predicting the Correctness Margin for Model Comprehension

The effectiveness of a model's self-reflection largely hinges on its "correctness margin," a metric quantifying its understanding of both the question and its context. Ideally, we would like to predict this margin through user prompts, thereby allowing the user to make an informed decision on whether to enable the model's self-reflection capability.

Nevertheless, our experiments indicate that current models struggle to self-assess their understanding reliably. Below, we outline our prompt design used for this experiment:

```
messages=[
    {"role": "system",
    "content": "You are a helpful
    assistant. Answer the question based
    on the context provided. Provide
    extremely concise answers with no
    explanation."},
    {"role": "user",
    "content": f"Context:
    {formatted_context}\n
```

```
        Question: {question}"},
        {"role": "assistant",
        "content": f"{response}"},
        {"role": "user",
        "content": "\nYou have just answered
        a question. Now, please evaluate
            ↪ your
        own comprehension of the question
            ↪ and
        answer provided. Rate your level of
        understanding on a scale from -5 to
            ↪ 5.
        A rating of 5 signifies extreme
        certainty that you understand the
        question, while a rating of -5
        indicates extreme uncertainty or
            ↪ lack
        of understanding."},
]
```

We tested this prompt structure on two sets of questions: one where all 10 model responses were incorrect, and another where all 10 were correct. If the model were capable of accurately evaluating its own comprehension, it should consistently rate its understanding at -5 for questions in the all-wrong dataset and 5 for those in the all-right dataset. However, after experimenting with 20 examples from each dataset, we found that the model consistently assigned high scores (typically 4 or 5) regardless of the dataset origin. Thus, reliable self-assessment remains an open challenge for current models.

G Scientific Artifacts

In this paper, we use the following artifacts:

- TruthfulQA (Lin et al., 2022) is a benchmark assessing a language model's ability to generate truthful answers for 817 diverse questions in 38 categories, requiring models to avoid false answers commonly found in human texts due to misconceptions or false beliefs. We use it for the preliminary studies on reflective thinking in LLMs. It is licensed under the Apache License, Version 2.0.

- HotpotQA (Yang et al., 2018) is a 113k question-answer dataset based on Wikipedia that requires multi-document reasoning, features diverse questions unconstrained by knowledge bases or schemas, provides sentence-level supporting facts for strong supervision and explanation, and introduces a new factoid comparison question type to evaluate QA systems' extraction and comparison abilities. We use it for evaluating reflective thinking in LLMs. It is distributed under a CC BY-SA 4.0 License.

- openai-python[7] (v0.27.8) provides convenient access to the OpenAI REST API from any Python 3.7+ application. We use it to access ChatGPT models. It is licensed under the Apache License, Version 2.0.

[7] https://github.com/openai/openai-python