Improving LLM Generations via Fine-Grained Self-Endorsement

Ante Wang[1,2,3], **Linfeng Song**[4]*, **Baolin Peng**[4], **Ye Tian**[4], **Lifeng Jin**[4], **Haitao Mi**[4],
Jinsong Su[1,2,3]*and **Dong Yu**[4]

[1]School of Informatics, Xiamen University, China
[2]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage
of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China
[3] Shanghai Artificial Intelligence Laboratory, China
[4]Tencent AI Lab, Bellevue, WA

wangante@stu.xmu.edu.cn, lfsong@global.tencent.com, jssu@xmu.edu.cn

Abstract

This work studies mitigating fact-conflicting hallucinations for large language model (LLM) at inference time. Particularly, we propose a self-endorsement framework that leverages the fine-grained fact-level comparisons across multiple sampled responses. Compared with prior ensemble methods (e.g., self-consistency (Wang et al., 2022; Chen et al., 2023)) that perform response-level selection, our approach can better alleviate hallucinations for knowledge-intensive tasks. Our approach can broadly benefit smaller and open-source LLMs as it mainly conducts simple content-based comparisons. Experiments on Biographies show that our method can effectively improve the factuality of generations with simple and intuitive prompts across different scales of LLMs. Besides, comprehensive analyses on TriviaQA and GSM8K demonstrate the potential of self-endorsement for broader application.[1]

1 Introduction

Recent Large Language Models (LLMs) such as LLaMA (Touvron et al., 2023) and Mixtral (Jiang et al., 2024) take billions of parameters and are trained on huge corpora of text documents with billions of tokens. As a result, they have demonstrated remarkable capabilities across various tasks such as longform generation, closed book QA and math reasoning. However, LLMs can still fail frequently on these knowledge-intensive and reasoning tasks where obviously incorrect facts or reasoning steps are generated. To address this issue, previous work has explored multiple orthogonal directions, such as introducing external knowledge and tool (Mallen et al., 2023; Peng et al., 2023; Wang et al., 2023c), continual supervised finetuning (Wu et al., 2023; Tian et al., 2023) and inference-time improvement (Dhuliawala et al., 2023; Chen et al., 2023) to reduce hallucination and improve reasoning capability. Among these research directions, inference-time improvement has recently gained popularity. The motivation behind may stem from various reasons: it can be used on black-box LLMs (e.g., no requirement on accessing the model weighs); it can work together with supervised finetuning by producing high-quality training data (a.k.a., self-distillation (Huang et al., 2022)).

Many prior approaches of inference-time improvement can be grouped into two main directions. The *ensemble* methods like self-consistency (Wang et al., 2022) and universal self-consistency (Chen et al., 2023) build upon traditional ensemble learning by picking the optimal prediction from multiple candidates sampled from the target LLM. Conversely, in the other direction, *self-refinement* methods such as chain-of-verification (Dhuliawala et al., 2023) and self-reflection (Madaan et al., 2023; Shinn et al., 2023) leverage the target LLM to refine its own predictions from varied perspectives. Comparatively, the ensemble methods can eliminate occasional hallucinations by looking into multiple peering samples. But, they may fail on longform generation tasks because the sampled candidates disagree with each other on too many places, making it difficult to pick the best prediction. More importantly, they cannot combine the merits from the peering samples. On the other hand, the self-refinement methods perform fine-grained refinement. But they rely on the assumption that the target LLM is strong enough to provide helpful critique for refinement, and thus most experiments on them are conducted on state-of-the-art close-source LLMs (e.g., GPT4 (Achiam et al., 2023)).

In this work, we follow the line of inference-time improvement to study how and when fine-grained cross-response validation (endorsement) can reduce hallucination and improve reasoning quality. Particularly, we propose a framework to im-

*Corresponding Authors

[1] Code is available at https://github.com/DeepLearnXMU/Self-Endorsement.

Findings of the Association for Computational Linguistics: ACL 2024, pages 8424–8436
August 11-16, 2024 ©2024 Association for Computational Linguistics

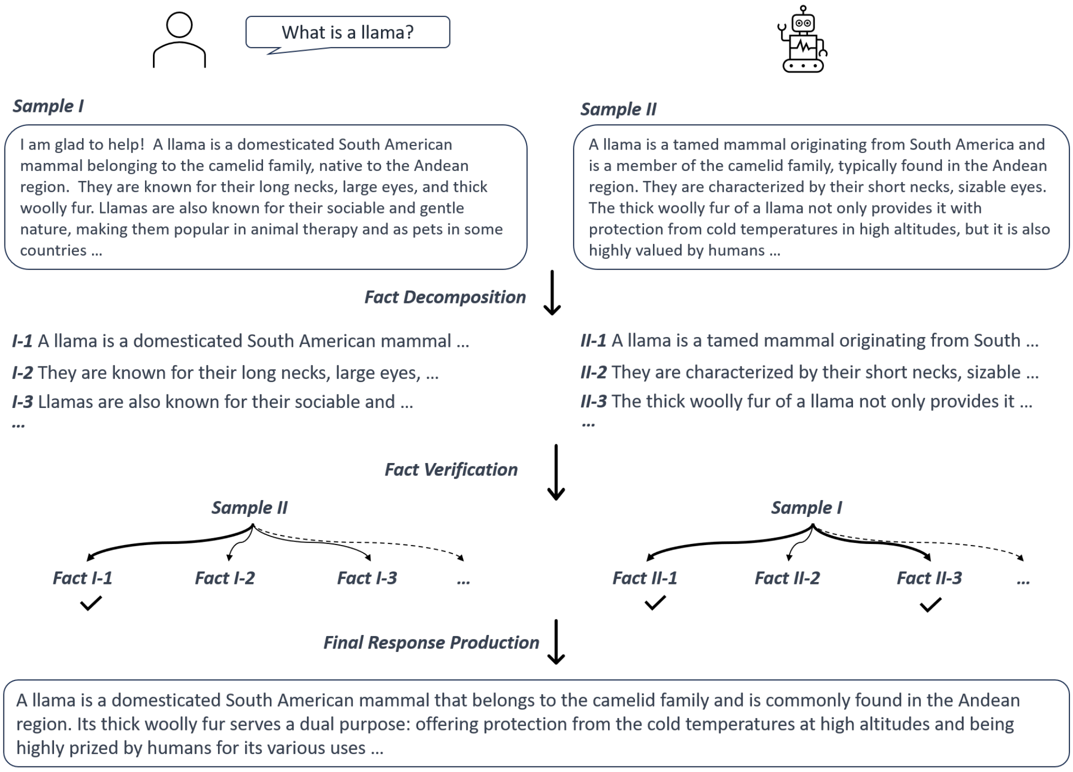

What is a llama?
Sample I
I am glad to help! A llama is a domesticated South American mammal belonging to the camelid family, native to the Andean region. They are known for their long necks, large eyes, and thick woolly fur. Llamas are also known for their sociable and gentle nature, making them popular in animal therapy and as pets in some countries ...
Sample II
A llama is a tamed mammal originating from South America and is a member of the camelid family, typically found in the Andean region. They are characterized by their short necks, sizable eyes. The thick woolly fur of a llama not only provides it with protection from cold temperatures in high altitudes, but it is also highly valued by humans ...
Fact Decomposition
I-1 A llama is a domesticated South American mammal ...
I-2 They are known for their long necks, large eyes, ...
I-3 Llamas are also known for their sociable and ...
...
II-1 A llama is a tamed mammal originating from South ...
II-2 They are characterized by their short necks, sizable ...
II-3 The thick woolly fur of a llama not only provides it ...
...
Fact Verification
Sample II
Fact I-1
Fact I-2
Fact I-3
...
Sample I
Fact II-1
Fact II-2
Fact II-3
...
Final Response Production
A llama is a domesticated South American mammal that belongs to the camelid family and is commonly found in the Andean region. Its thick woolly fur serves a dual purpose: offering protection from the cold temperatures at high altitudes and being highly prized by humans for its various uses ...

Figure 1: The example framework of self-endorsement, where only two sampled candidates are leveraged.

prove LLM predictions by leveraging fine-grained cross-response endorsements. As shown in Figure 1, we begin by generating multiple samples from the target LLM. Next, we extract facts from each sample and prompt the LLM to verify the endorsement of each fact by cross-referencing with the other samples. An endorsement score is then assigned to each fact based on its level of approval. Finally, to produce the final response, we either select the sample with the most reliable facts or regenerate a new one by incorporating the facts with high endorsement scores as supplementary inputs to the LLM. Without complex instructions, the LLM is only required to conduct two tasks: 1) check whether a fact is consistent with the knowledge in another response at a time; 2) generate a new response given additional high-quality facts as inputs. Both tasks are fairly simple, thus we believe (and our experiments show that) our method can be broadly helpful to various open-source LLMs of different capacities.

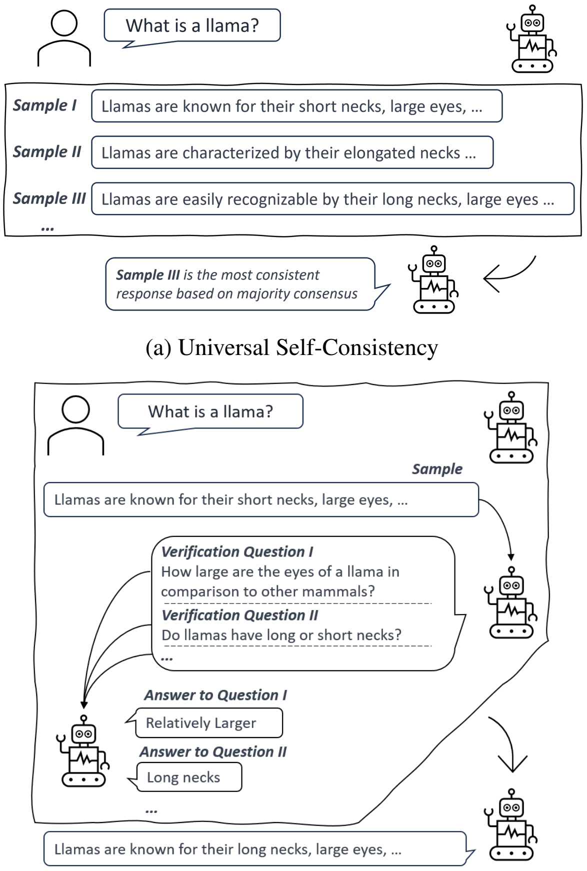
We mainly conduct experiments to examine the level of fact-conflicting hallucinations in model predictions. Specifically, we evaluate on Biographies (Min et al., 2023), a benchmark of longform generation, and TriviaQA (Joshi et al., 2017), a popular dataset on generative QA. Results on popular open-source models, such as LLaMA2 (Touvron et al., 2023) and Mixtral (Jiang et al., 2024), show that our method greatly reduces hallucination by a large margin. Details analyses suggest that our method can better select reliable fine-grained facts across various model sizes. Study on GSM8K (Cobbe et al., 2021), a benchmark of math word problems, further validates the promise of self-endorsement for more pervasive use.

2 Baselines

We take (universal) self-consistency (Wang et al., 2022; Chen et al., 2023) and chain-of-verification (Dhuliawala et al., 2023) as the baseline for comparison. They are two popular methods of inference-time improvement based on ensemble learning and self-refinement, respectively.

2.1 (Universal) Self-Consistency

Self-consistency (SC) is a majority-voting-based ensemble approach designed for reasoning tasks. Specifically, it first samples multiple reasoning paths and their corresponding answers from the LLM, e.g., (r_i, a_i), where $r_i \Rightarrow a_i$. It then selects the most consistent answer via taking a majority vote over a_i, i.e., $max_{\mathrm{a}} \sum_i \mathbb{1}(a_i = \mathrm{a})$. With chain-of-thought prompting (CoT), it has demonstrated


What is a llama?
Sample I
Llamas are known for their short necks, large eyes, ...
Sample II
Llamas are characterized by their elongated necks ...
Sample III
Llamas are easily recognizable by their long necks, large eyes ...
...
Sample III is the most consistent response based on majority consensus
What is a llama?
Sample
Llamas are known for their short necks, large eyes, ...
Verification Question I
How large are the eyes of a llama in comparison to other mammals?
Verification Question II
Do llamas have long or short necks?
...
Answer to Question I
Relatively Larger
Answer to Question II
Long necks
...
Llamas are known for their long necks, large eyes, ...

(a) Universal Self-Consistency

(b) Chain-of-Verification

Figure 2: Two main baselines in this work.

remarkable performance gains on complex reasoning tasks. However, self-consistency can only be applied to tasks where the final answer can be aggregated via exact match (e.g., question answering and math word problems).

To support broader applications, universal self-consistency (USC) extends self-consistency by taking the LLM itself (instead of majority voting) to select the final response from the samples it generated. Particularly as shown in Figure 2a, the LLM is first asked to sample multiple candidates, it then consumes all these candidates to pick one as the final response. To achieve precise final answer selection, USC may require that the LLM possesses robust critical analysis capabilities.

2.2 Chain-of-Verification

Different from ensemble-based SC / USC, chain-of-verification (CoVe) refines factual errors in one response and then regenerates a new one by the LLM itself. As shown in Figure 2b, the LLM is asked to first (I) draft an initial sample; then (II) plan verification questions to fact-check its draft; (III) answer those questions independently; and (IV) generate its final verified response.

The core motivation of CoVe is that LLMs tend to provide more accurate facts to simple questions (e.g., the verification questions) than complex questions (e.g., the original question). Hence it can improve the factuality of the overall response.

3 Self-Endorsement

As shown in Figure 1, our self-endorsement framework interacts with an LLM by taking the following steps given a user query $\mathcal{X}$:
(1) *Candidate Sampling*: It asks the LLM to sample N candidate responses $Y_1, Y_2, ..., Y_N$.
(2) *Fact Decomposition*: It breaks down each candidate Y_i into facts $f_1^i, f_2^i, ..., f_{N_{Y_i}}^i$, where N_{Y_i} is the number of facts in Y_i.
(3) *Fact Verification*: It verifies each fact f_j^i via calculating its endorsement scores against other candidates $\{Y_k \mid k \neq i\}$. We also explore context pruning, which eliminates unrelated content in candidates for verification.
(4) *Final Response Production*: Produce a final response via selection or regeneration. Specifically, we either select the response with facts having the highest endorsement scores as the final response or ask the LLM to regenerate a new one Y given the set of selected facts $\mathcal{Z}$ from different candidates.

3.1 Candidate Sampling

We follow the common practice of sampling N responses via nucleus sampling. Each sampling process is denoted as $Y_i \sim \texttt{LLM}(\mathcal{X})$.

3.2 Fact Decomposition

Following exiting work (Gao et al., 2022; Liu et al., 2023), we consider a fact as a statement about some factual knowledge. There are many ways to conduct fact decomposition. We first adopt a naive method used by some previous work (Liu et al., 2023; Manakul et al., 2023), which takes each sentence in a response as a fact. However, it fails to consider the situations that some sentences can contain multiple independent facts (Liu et al., 2023) or do not contain any fact. Therefore, we also study prompting the LLM itself to extract facts from its responses. This process is denoted as $f_1^i, f_2^i, ..., f_{N_{Y_i}}^i = \texttt{LLM}(Y_i, P_D)$, where P_D is the corresponding LLM instruction shown below:

> *List all non-repeated facts from the text below in numerical order. Each fact should be a self-contained sentence:* Y_i

We observe that the LLM-prompting method can effectively eliminate statements without factual knowledge and break down complex sentences into multiple pieces of facts.

3.3 Fact Verification by Self-Endorsement

We verify each fact via its endorsement score: the degree of the fact being consistent with the content in other sampled responses. There are multiple ways to compare two pieces of text, such as querying the LLM or calling a sentence encoder (e.g. SimCSE (Gao et al., 2021)). For simplicity and to minimize the effect of extra supervision, we choose to query the LLM via prompting.

Formally, for a fact f_j^i from response Y_i, we feed f_j^i and another response Y_k ($k \neq i$) to the LLM with prompt P_V to determine whether Y_k endorses f_j^i. Then, we define the endorsement score of f_j^i as

$$g(f_j^i) = \frac{1}{N-1} \sum_{k \neq i} \mathbb{I}[\texttt{LLM}(f_j^i, Y_k, P_V) \text{ is } \textit{true}].$$

The prompt P_V is simply defined as

Take the following as truth: Y_k
Then the following statement: "f_j^i" is true, false, or inconclusive?

In many situations, especially for longform generation, most facts in Y_k can be irrelevant to f_j^i. Therefore, we propose to further prune the unnecessary context and only keep the most related parts to speed up inference. Particularly, we select top-K similar facts to f_j^i from each Y_k using the BM25 algorithm. Then, we concatenate the K selected facts (denoted as Y_k') to verify f_j^i.

Generally, the endorsement score reflects the level of confidence from the LLM to a piece of fact. Therefore, facts with higher endorsement scores have higher chances to be faithful.

3.4 Selection / Regeneration for Final Response Production

Selection After the above steps, a simple option is to select one from the sampled candidates as the final response Y. For each candidate Y_i, we average the endorsement scores of its facts (i.e., $\texttt{Avg}(g(f_1^i), ...)$) and select the one with the highest average score as the final response. However, this does not fully exploit the potential of our framework due to the following reasons: (1) There can still be factual errors in the selected response. (2) Helpful and complementary facts in other responses are not efficiently leveraged.

Regeneration We propose another option that prompts the LLM to regenerate the final response Y with selected facts ($\mathcal{Z}$) from all samples: $Y \sim \texttt{LLM}(\mathcal{X}, \mathcal{Z}, P_G)$, where prompt P_G is defined as

Knowledge from other sources: $\mathcal{Z}$
Given the materials above, answer the question: $\mathcal{X}$

To select useful facts, we first discard the facts whose endorsement scores do not exceed a threshold α (i.e., $g(f_j^i) \leq \alpha$). Though this can effectively prune low-quality facts, there can still be facts of redundant content. We then adopt a K-means algorithm that takes bag-of-words features as the representation for each fact and groups the facts into $\mathcal{C}$ clusters. Lastly, we select the fact closest to the centroid for each cluster to form the selected fact set $\mathcal{Z}$ that contains $|\mathcal{C}|$ facts.

4 Experiments

4.1 Setup

Datasets We mainly conduct experiments on Biographies (Min et al., 2023) and TriviaQA (Joshi et al., 2017). Biographies is a popular benchmark focusing on knowledge-intensive longform text generation. It contains 183 person entities used to prompt LLMs about their biographies with the query *"Tell me a bio of <entity>"*. TriviaQA (Joshi et al., 2017) is a popular open-domain question-answering benchmark. We do not add restrictions (e.g., early stopping or instructing the LLM to only generate the answer) to encourage the LLM to generate explanations and relevant knowledge in addition to the answer. For evaluation, we report answer recall (*Ans. Rec.*) in addition to *Fact Acc.* and *#Fact*. For math reasoning, we test self-endorsement on GSM8K (Cobbe et al., 2021). More details about both datasets are introduced later in this section.

Evaluation For Biographies, we follow Min et al. (2023) to evaluate the accuracy of decomposed facts (*Fact Acc.*) using their released inst-LLaMA-7B model together with the Wikipedia dump from 2023/04/01 as judge. Particularly, the correctness of each fact is evaluated by inst-LLaMA-7B that takes the top 5 passages retrieved from the wiki page of the topic entity as extra evidence. Though inst-LLaMA-7B is much smaller than the start-of-the-art LLMs such as ChatGPT, Min et al. (2023) has shown that inst-LLaMA-7B can always give consistent judging decisions with ChatGPT. In addition to *Fact Acc.*, we also report the number of facts (*#Fact*), because good responses should contain a decent number of facts of high accuracy.

For TriviaQA, we follow standard practice to also report answer recall (*Ans. Rec.*) in addition to fact accuracy and the number of facts. Answer recall measures if the target answer is contained in the generated response. For GSM8K, we report the quality of the intermediate reasoning steps using GPT4 as judge (*GPT4 (Y)* and *GPT4 (N)*) in addition to the accuracy of the final answer (*Acc.*). More details on the quality of the intermediate steps are introduced in the corresponding section.

Settings and Hyperparameters We conduct experiments based on LLaMA2-7B-Chat, LLaMA2-70B-Chat (Touvron et al., 2023) and Mixtral-8x7B-Inst (Jiang et al., 2024) for Biographies and TriviaQA. Only Mixtral-8x7B-Inst is adopted for GSM8K due to its stronger math capabilities.

For our approach, we use nucleus sampling with a temperature of 1.0 when generating responses and use greedy decoding otherwise. We prompt the target LLM to extract facts for Biographies and TriviaQA and directly take each sentence in a response as a fact for GSM8K. We empirically set candidate number N (§3.1), the number of kept context facts K (§3.3), and fact-filtering threshold α (§3.4) as 10 / 10, 3 / 3, 1.0 / 0.8 for LLaMA2-7B-Chat / LLaMA2-70B-Chat. The K-means cluster number $\mathcal{C}$ is dynamically decided by the average number of facts across the N candidate responses. We also conduct careful analyses on the effects of these hyperparameters.

Baselines One obvious baseline is simply calling LLM to sample a response. We report the average numbers from N sampled responses to alleviate the randomness of the sampling process. In addition, we take the following baselines for a better understanding of our approach:

- *Refine*: Considering the power of LLMs, an LLM might be able to correct its own errors given a second chance. This baseline is set to quantify the gain from this effect.

- *(Universal) Self-Consistency (SC / USC)*: They are implemented as mentioned in §2.1.

- *Chain-of-Verification (CoVe)*: Its implementation follows the description in §2.2.

Model	Fact Acc.	#Fact
LLaMA2-7B-Chat	53.2	16.8
+refine	52.6	15.7
+USC	53.5	15.9
+CoVe	54.8	9.8
self-endorsement		
+select	58.2**	15.9
+select w/ pruning	59.6**	15.2
+regenerate	**67.7****	14.9
+regenerate w/ pruning	65.7**	14.6
LLaMA2-70B-Chat	63.1	20.0
+refine	64.9*	20.2
+USC	61.6	20.4
+CoVe	64.0	16.5
self-endorsement		
+select	66.5**	19.4
+select w/ pruning	67.7**	18.8
+regenerate	**73.1****	18.3
+regenerate w/ pruning	73.0**	17.9

Table 1: Test results on Biographies. Results using Mixtral-8x7B-Inst is in Table 5 of Appendix due to limited space. We conduct bootstrap resampling for significant test. *, ** denote significantly better results over the base LLM (the first line in each group) with significance level $p < 0.05$ and $p < 0.01$, respectively.

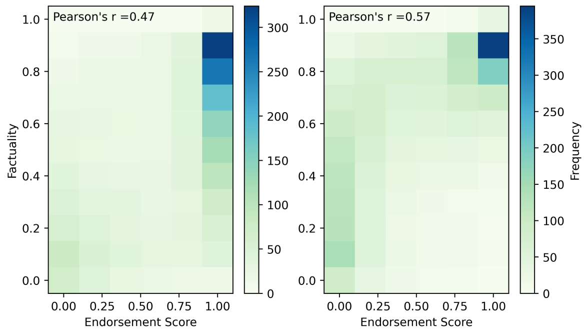
4.2 Results and Analyses

Self-endorsement Helps Improving Factuality As shown in Table 1, none of the baselines (*+refine*, *+USC* and *+CoVe*) can significantly improve over the 7B and 70B LLaMA2-Chat model regarding *Fact Acc.* In contrast, self-endorsement gives significant improvements over baselines no matter whether the final response is selected or regenerated and whether context pruning is used or not.

Among those baselines, only *CoVe* can slightly improve *Fact Acc.*, but it obviously decreases the *#Fact*, which is also observed in Dhuliawala et al. (2023). *Refine* only benefits LLaMA2-70B-Chat, while the gain is still much inferior to our self-endorsement approaches based on self-selected high-quality facts. The results of *Refine* also indicate that naive self-refinement demands strong capabilities of the LLM.

For our methods, because regeneration can include reliable facts from all candidates and discard incorrect facts, thus it consistently produces better responses than selection. Using context pruning or not gives a minor performance change regarding *Fact Acc.* We will provide more analyses in later experiments.

Endorsement Score Correlates with Factuality Since endorsement scores play a crucial role in the success of our approaches, we further investigate



(a) LLaMA2-7B-Chat

(b) LLaMA2-70B-Chat

Figure 3: Statistical correlation between endorsement scores and factuality scores.

K	Fact Acc.	#Fact
1	62.5	15.1
3	65.7	14.6
5	66.8	14.7
ALL	**67.7**	14.9
1	72.4	18.1
3	73.0	17.9
5	**73.2**	18.2
ALL	73.1	18.3

Table 2: Performances on LLaMA2-7B-Chat (up) and LLaMA2-70B-Chat (down) when using K facts from other responses to calculate the endorsement score for target facts.

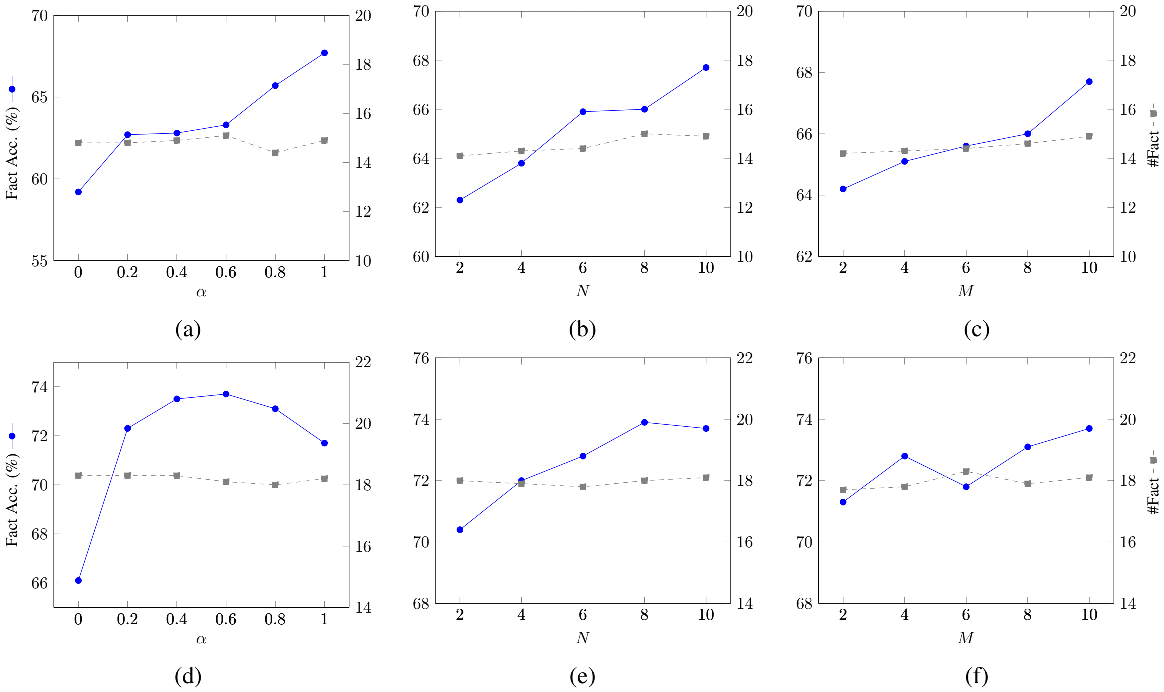
how endorsement scores are correlated with the actual factuality. To this end, we use inst-LLaMA-7B with Wikipedia dump to calculate the factuality score for each piece of fact. Figure 3 presents the correlation between endorse scores and factuality scores. Results on both models show clear positive relationships between endorsement scores and factuality. LLaMA2-70B-Chat gives a stronger correlation because of its stronger ability over LLaMA2-7B-Chat. Especially, LLaMA2-7B-Chat tends to give higher endorsement scores to certain incorrect facts erroneously.

How the Quality of Selected Facts Affect Final Responses? Since threshold α decides the quality of selected facts for regeneration, here we try several values of α and visualize the corresponding final-response quality in Figure 4a and 4d. We observe that ranging α from 0 to 1 keeps benefiting LLaMA2-7B-Chat but the performance on LLaMA2-70B-Chat is increased first and then decreased. After a closer look, we find that a high α may limit the quantity and diversity of selected facts, which may hurt the regeneration quality. For example, when $\alpha = 1$, only an average number of 11.3 facts are selected under LLaMA2-70B-Chat, while the number is 16.7 for LLaMA2-7B-Chat. Besides, we observe a decent performance increase with $\alpha \geq 0.2$, showing the effectiveness of our approach on alleviating the side-effect of low-quality facts by removing them.

How Candidate Number Affects Final Responses? Intuitively, increasing the candidate number N can help to provide more high-quality facts and each fact can also be better verified with more samples. As shown in Figure 4b and 4e, the performances of both 7B and 70B models generally get improved when increasing N, and the number of facts in regenerated responses remains stable. For LLaMA2-7B-Chat, more improvements can be expected when N is further increased. However, this will also bring more computational costs that can be impractical. In contrast, LLaMA2-70B-Chat is less sensitive, showing that a small N is enough for stronger LLMs. Encouragingly, we also observe that our models can significantly outperform baselines with limited samples (70.4 vs. 63.1 when $N = 2$ on LLaMA2-70B-Chat). This suggests the robustness of our method in some extreme cases.

Effect on Selecting Facts from Fewer Candidates for Regeneration We further analyze the effect of selecting facts from fewer number (denoted as M and $M < N$) of candidates. Note that these facts from the M candidates can still take all N candidates to calculate their endorsement scores. Results are shown in Figure 4c and 4f. We again observe positive effects when increasing M, because the final responses can directly consult more provided input facts. Besides, by comparing the results in Figure 4b and 4c (also Figure 4e vs 4f), we find that the latter performs better when the candidate number is small (e.g., 71.3 vs. 70.4 when both $N = 2$ and $M = 2$ on LLaMA2-70B-Chat). This indicates that a fact can be better verified when more candidates are available for calculating endorsement scores.

How Context Pruning Affects Final Responses? Context pruning aims to eliminate unnecessary context when calculating the endorsement score for each fact, while it may hurt the accuracy of fact selection and overall performance when too much context is pruned. As shown in Table 2 (up), LLaMA2-7B-Chat is largely influenced by K, and its performance stably improves when K increases.



(a) (b) (c)

(d) (e) (f)

Figure 4: Hyperparameter analyses on LLaMA-7B-Chat (up) and LLaMA-70B-Chat (down). We present different choices of α, N and M and their effects on *Fact Acc.* and *#Fact*.

Conversely, though growing *Fact Acc.* scores are observed as well for LLaMA2-70B-Chat (Table 2 (down)), the growth rate is mild (e.g., 72.4 → 73.2). This is consistent with the comparison on both candidate number N (Figure 4b vs 4e) and candidate number for fact selection M (Figure 4c vs 4f). For both 7B and 70B models, we observe *Fact Acc.* numbers that are close to when no context pruning is used. Thus, context pruning is useful overall, especially considering that it can save 50% of computation cost when $K = 5$ according to statistics. Note that we only use the vanilla BM25 algorithm for selecting related facts. We leave exploring better sentence matching algorithms in future work.

Evaluation Results on Question Answering Results are shown in Table 3. Our method again effectively improves the *Fact Acc.*, which is consistent with our observations on Biographies. The improvements regarding *Ans. Rec.* are limited. It is because LLMs have already provided more accurate exact answers to target questions (Dhuliawala et al., 2023) but tend to ignore other facts in the responses. Besides, *regeneration* gives fewer improvements over *selection* on this dataset, which can be due to the limited fact numbers in short-text generation thus *selection* is also easier to select a good one from enough candidates.

Model	Fact Acc.	Ans. Rec.	#Fact
LLaMA2-7B-Chat	57.4	70.0	4.8
+USC	57.6	69.0	4.8
+CoVe	53.7	**71.2**	4.3
self-endorsement			
+select	63.4**	70.2	4.4
+select w/ pruning	63.8**	69.5	4.3
+regenerate	**65.0****	70.7	4.7
+regenerate w/ pruning	64.0**	70.8	4.4
LLaMA2-70B-Chat	65.1	84.1	5.0
+USC	65.0	83.1	5.0
+CoVe	58.9	83.1	5.4
self-endorsement			
+select	69.7**	83.8	4.8
+select w/ pruning	70.2**	84.2	4.7
+regenerate	**71.7****	**85.3***	5.2
+regenerate w/ pruning	70.7**	85.0*	5.2

Table 3: Test results on TriviaQA. Results using Mixtral-8x7B-Inst is in Table 6 of Appendix due to limited space.

Extensive Experiments on GSM8K In addition to knowledge-intensive tasks, we also briefly explore self-endorsement on reasoning tasks, choosing GSM8K (Cobbe et al., 2021), a popular math benchmark, as the testbed. Here we focus more on the quality of intermediate reasoning steps in addition to the final-answer accuracy (*Acc.*). Particularly, we divide the reasoning steps into two groups (*Yes*/*No*) based on whether their corresponding predicted answers are correct or not. We then prompt-

Model	Acc.	GPT4 (Y)	GPT4 (N)
Mixtral-8×7B-Inst	68.4	9.87	3.65
+USC	71.6*	9.86	3.90*
+CoVe	56.0	–	–
+SC	80.3**	9.87	3.96**
+select	**80.8****	9.87	**4.08****

Table 4: Test results on GSM8K. We do not report *CoVe* results on GPT4 because its answers usually do not contain complete rationales.

ing gpt-4-0613 with the instruction[2] from the MT-bench (Zheng et al., 2023) to measure the quality of each group (*GPT4 (Y)* / *GPT4 (N)*).

As shown in Table 4, both *USC* and *SC* help improve *Acc.* while *SC* performs significantly better. This is because *SC*, which conducts majority voting on final answers, is more aligned with *Acc.* *CoVe* even severely hurt model performance. This is because the augmented questions occasionally inquire about irrelevant topics, which disturb the main reasoning procedure.

Regarding the intermediate steps, there is a large performance gap between both groups (*Yes*/*No*). Thus, how to further improve the group of incorrect final answers has become critical. Our method reports a slightly better result than *SC* on *Acc*, and the gap on *GPT4 (N)* is even more (0.12 over 10). This indicates that our method indeed helps select relatively better rationales even though the final answers are incorrect, validating the effectiveness of our method from another aspect.

5 Related Work

5.1 Inference-time Hallucination Mitigation

Researchers have explored mitigating LLM hallucinations at both training and inference time. Compared with training-time mitigation approaches (Lee et al., 2022; Lightman et al., 2023; Tian et al., 2023; Lan et al., 2023; Wu et al., 2023; Zhang et al., 2024), inference-time improvement is gaining popularity because it can be more cost-effective and controllable (Zhang et al., 2023).

Many studies (Shi et al., 2023; Wang et al., 2023a,b; Huang et al., 2024) have resorted to external knowledge for improving factuality of LLMs by first retrieving relevant knowledge from databases or tools (e.g., search engines) before providing LLMs for prediction. In contrast, we delve into another research line (Lee et al., 2022; Chen et al., 2023; Dhuliawala et al., 2023; Li et al., 2023; Chuang et al., 2023; Das et al., 2024) that mitigates hallucinations exclusively through the utilization of the LLM itself, without any external assistance. This can be crucial in situations where external knowledge sources are unavailable.

Except for the two baselines *USC* and *CoVe* we introduced previously, Lee et al. (2022) proposed *factual-nucleus sampling* that balances diversity and factuality by dynamically adjusting the hyperparameters of sampling when decoding. Li et al. (2023) introduced *Inference-Time Intervention (ITI)* that shifts model activations along truth-correlated directions after identifying attention heads with high linear probing accuracy for truthfulness. Chuang et al. (2023) found that factual information is encoded in distinct layers, thus they contrasted the generation probabilities from different layers of LLMs. Among these studies, our approach is most related to *USC* involving checking the consistency across sampled candidates but is conducted at the fact level.

5.2 Black-box Hallucination Detection

Detecting hallucinations during inference is usually based on uncertainty estimation. Current work can be categorized into three types (Zhang et al., 2023): logit-based (Guo et al., 2017), verbalize-based (Xiong et al., 2023), and consistency-based (Manakul et al., 2023; Mündler et al., 2023).

This work is most relevant to the consistency-based approach, which operates on the assumption that LLMs are likely to provide logically inconsistent responses for the same question when they are indecisive and hallucinating facts (Zhang et al., 2023). For instance, SelfCheckGPT (Manakul et al., 2023) explored several methods, such as BERTScore (Zhang et al., 2019), to check informational consistency between sampled responses. Mündler et al. (2023) utilized an additional LLM to detect incorrect facts by checking whether there is a contradiction between two responses given the same context. Our method shares similarities with these approaches in terms of checking consistency among sampled responses. Nonetheless, our endorsement scores are calculated at a finer level (fact vs fact). More importantly, we prioritize improving the quality of final responses after detecting hallucinations.

[2] See Figure 9 in Appendix.

6 Conclusion

In this paper, we present self-endorsement, a framework that alleviates hallucinations and improves reasoning capability solely by the LLM itself. Particularly, we first perform fine-grained fact-level comparisons among multiple sampled candidates to identify reliable facts. Then, we produce the final response by either selecting from candidates or regenerating based on these facts. We evaluate our approach on popular benchmarks including Biographies for the longform generation, TriviaQA for open-domain question answering, and GSM8K for mathematical multi-step reasoning. Results show that self-endorsement can significantly benefit small or open-source LLMs without intricate instructions compared with previous approaches.

Limitations

The main limitation of self-endorsement lies in the computation cost incurred at the fact verification phase. The cost escalates dramatically when using more candidates for collecting verified facts. In this work, we have demonstrated the trade-off between candidate numbers and final performance: limited candidate numbers can still help improve factuality and larger models exhibit less sensitivity to hyperparameter selection. Future studies can also explore quantization (Jacob et al., 2018) or distilling knowledge into a smaller model (Hinton et al., 2015) to improve computational efficiency further. Another limitation is that our method is fully based on prompting. Given the sensitivity of LLMs to input prompts, the choice of prompts can impact final performance. Moreover, a single prompt may not consistently yield optimal results across diverse tasks or models. Techniques used for prompt searching can help solve this problem (Yang et al., 2023). We leave this as future work.

Self-endorsement follows the line of research that reduces LLM hallucinations using only the knowledge from the LLM. Thus, its performance is inherently limited by the capacity of the target LLM. For example, statements with high endorsement scores may still contain factual inaccuracies due to outdated or noisy knowledge embedded inside the model parameters. Since the LLM does not possess accurate information about these facts, it is impossible to prevent hallucinations without incorporating external knowledge, as discussed in (Simhi et al., 2024).

Acknowledgments

The project was supported by National Key R&D Program of China (No.2022ZD0160501), National Natural Science Foundation of China (No.62276219), and the Public Technology Service Platform Project of Xiamen (No.3502Z20231043). We also thank the reviewers for their insightful comments.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Souvik Das, Lifeng Jin, Linfeng Song, Haitao Mi, Baolin Peng, and Dong Yu. 2024. Entropy guided extrapolative decoding to improve factuality in large language models. *arXiv preprint arXiv:2404.09338*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2022. Attributed text generation via post-hoc research and revision. *arXiv preprint arXiv:2210.08726*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jianheng Huang, Ante Wang, Linfeng Gao, Linfeng Song, and Jinsong Su. 2024. Response enhanced semi-supervised dialogue query generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18307–18315.

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*.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Zhibin Lan, Wei Li, Jinsong Su, Xinyan Xiao, Jiachen Liu, Wenhao Wu, and Yajuan Lyu. 2023. Factgen: Faithful text generation by factuality-aware pretraining and contrastive ranking fine-tuning. *Journal of Artificial Intelligence Research*, 76:1281–1303.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. 2024. Constructing benchmarks and interventions for combating hallucinations in llms.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ante Wang, Linfeng Song, Qi Liu, Haitao Mi, Longyue Wang, Zhaopeng Tu, Jinsong Su, and Dong Yu. 2023a. Search-engine-augmented dialogue response generation with cheaply supervised query production. *Artificial Intelligence*, 319:103874.

Ante Wang, Linfeng Song, Ge Xu, and Jinsong Su. 2023b. Domain adaptation for conversational query production with the rag model feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9129–9141.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Y Wu, and Zhifang Sui. 2023c. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *CoRR, abs/2312.08935*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. 2023. Eva-kellm: A new benchmark for evaluating knowledge editing of llms. *arXiv preprint arXiv:2308.09954*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-contrast: Better reflection through inconsistent solving perspectives. *arXiv preprint arXiv:2401.02009*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Examples of Self-Endorsement Prompts

We present an example of our implementation on LLaMA2-70B-Chat, sourced from TriviaQA, as depicted in Figures 5, 6, 7, and 8. The responses of this dataset are concise and thus more suitable to display.

Model	Fact Acc.	#Fact
Mixtral-8×7B-Inst	76.3	18.2
+USC	76.4	18.1
+CoVe	67.2	11.6
self-endorsement		
+select	79.3	18.0
+regenerate	**84.1**	17.3

Table 5: Test results on Biographies.

B Experiments on Mixtral-8×7B-Inst

We also report results on Mixtral-8×7B-Inst for Biographies and TriviaQA in this section. Table 5 and Table 6 again demonstrate the effectiveness of our approach in defeating hallucinations.

Model	Fact Acc.	Ans. Rec.	#Fact
Mixtral-8×7B-Inst	65.5	85.4	4.3
+USC	65.4	**87.7**	4.3
+CoVe	65.4	87.3	6.5
self-endorsement			
+select	67.5	87.2	4.0
+regenerate	**68.3**	87.3	5.5

Table 6: Test results on TriviaQA.

What colour cat is a Russian Blue?

Sample 0: A Russian Blue is a breed of domesticated cat that is known for its thick, plush coat and its distinctive blue-gray color. The coat can range in color from a light blue-gray to a dark charcoal-gray, and may have a slight silver or black sheen to it. The exact color of a Russian Blue can vary depending on the individual cat and its breeding, but it is generally a consistent shade of blue-gray. Some Russian Blues may have a slight tabby pattern, known as "lapsed" or "Classic" pattern, which is a result of the interaction between the blue-gray color and the agouti gene. Overall, the Russian Blue is a beautiful and striking breed with a unique and recognizable coat color.

Sample 1: A Russian Blue is a breed of domestic cat that is known for its blue-gray coat. The exact color of a Russian Blue can vary depending on the individual cat and its breeding, but it is generally described as a cool, dark blue-gray color with a silver or frosty sheen. The fur may have a slight brown or yellow tint to it, but overall, the overall appearance is a sleek, silvery blue. The Russian Blue is a popular breed of cat known for its intelligence, affectionate personality, and striking appearance.

Sample 2: A Russian Blue is a breed of domesticated cat known for its thick, plush coat and silver-gray color. The coat of a Russian Blue is short in length, dense, and smooth, with a silver-gray undercoat and a darker, charcoal-gray overcoat. The color of a Russian Blue can range from a light silver gray to a dark charcoal gray, with a distinctive bluish tint. The exact color of a Russian Blue can vary depending on the individual cat and its breeding, but overall, they are known for their striking, elegant appearance and their signature silver-gray coat.

…

Figure 5: Step 1 - candidate sampling. We only display 3 candidate samples here and the input prompt is highlighted in blue.

List all non-repeated facts from the text below in numerical order. Each fact should be a self-contained sentence.

A Russian Blue is a breed of domesticated cat that is known for its thick, plush coat and its distinctive blue-gray color. The coat can range in color from a light blue-gray to a dark charcoal-gray, and may have a slight silver or black sheen to it. The exact color of a Russian Blue can vary depending on the individual cat and its breeding, but it is generally a consistent shade of blue-gray. Some Russian Blues may have a slight tabby pattern, known as "lapsed" or "Classic" pattern, which is a result of the interaction between the blue-gray color and the agouti gene. Overall, the Russian Blue is a beautiful and striking breed with a unique and recognizable coat color.

Facts:
1. A Russian Blue is a breed of domesticated cat that is known for its thick, plush coat and its distinctive blue-gray color.
2. The coat can range in color from a light blue-gray to a dark charcoal-gray, and may have a slight silver or black sheen to it.
3. The exact color of a Russian Blue can vary depending on the individual cat and its breeding, but it is generally a consistent shade of blue-gray.
4. Some Russian Blues may have a slight tabby pattern, known as "lapsed" or "Classic" pattern, which is a result of the interaction between the blue-gray color and the agouti gene.
5. The Russian Blue is a beautiful and striking breed with a unique and recognizable coat color.

Figure 6: Step 2 - fact decomposition. We take sample 0 in Figure 5 as an example.

Take the following as truth:
- A Russian Blue is a breed of domestic cat that is known for its blue-gray coat.
- The exact color of a Russian Blue can vary depending on the individual cat and its breeding.
- The Russian Blue is a popular breed of cat known for its intelligence.

Then the following statement: "The coat can range in color from a light blue-gray to a dark charcoal-gray, and may have a slight silver or black sheen to it." is true, false, or inconclusive?

Answer: True

Take the following as truth:
- A Russian Blue is a breed of domesticated cat that is known for its thick, plush coat and its distinctive blue-gray color.
- The Russian Blue is a beautiful and striking breed with a unique and recognizable coat color.
- The exact color of a Russian Blue can vary depending on the individual cat and its breeding, but it is generally a consistent shade of blue-gray.

Then the following statement: "The Russian Blue is a popular breed of cat known for its intelligence." is true, false, or inconclusive?

Answer: Inconclusive

Take the following as truth:
- A Russian Blue is a breed of domesticated cat that is known for its thick, plush coat and its distinctive blue-gray color.
- The Russian Blue is a beautiful and striking breed with a unique and recognizable coat color.
- The exact color of a Russian Blue can vary depending on the individual cat and its breeding, but it is generally a consistent shade of blue-gray.

Then the following statement: "A Russian Blue is a breed of domesticated cat known for its thick, plush coat and silver-gray color." is true, false, or inconclusive?

Answer: False

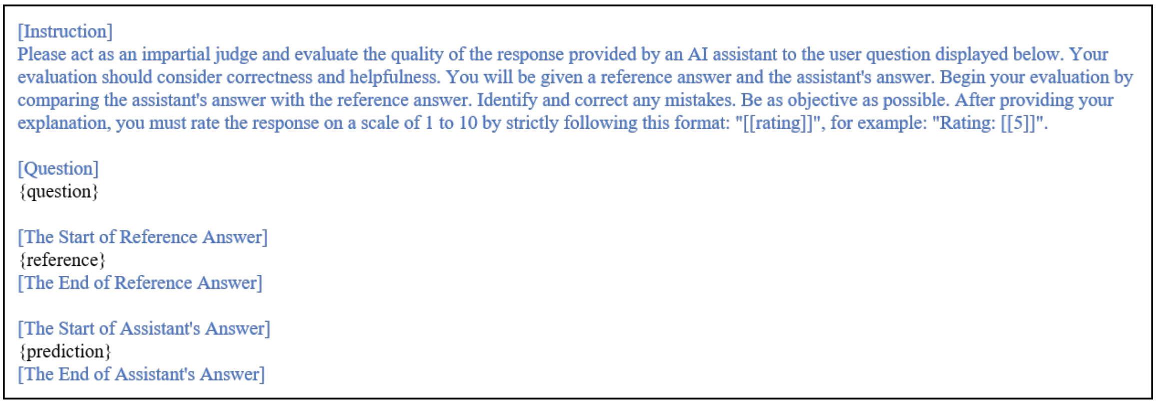
Figure 7: Step 3 - fact verification. We display 3 examples with different classification results.

Knowledge from other sources:
- A Russian Blue is a breed of domestic cat that is known for its blue-gray coat.
- The coat can range in color from a light blue-gray to a dark charcoal-gray, and may have a slight silver or black sheen to it.
- The exact color of a Russian Blue can vary depending on the individual cat and its breeding, but it is generally a consistent shade of blue-gray.
- Some Russian Blues may have a slight tabby pattern, known as \"lapsed\" or \"Classic\" pattern, which is a result of the interaction between the blue-gray color and the agouti gene.
- The Russian Blue is a beautiful and striking breed with a unique and recognizable coat color.
- The Russian Blue is a popular breed of cat known for its striking appearance.
- The color of a Russian Blue cat is actually a result of a genetic mutation that affects the production of the pigment melanin, which gives the coat its distinctive blue-gray appearance.

Answer the question below and then elaborate using the provided information:
What colour cat is a Russian Blue?

Final Response: A Russian Blue is a breed of domestic cat known for its blue-gray coat. The exact color can range from a light blue-gray to a dark charcoal-gray, with a slight silver or black sheen, and some individuals may have a slight tabby pattern. The unique and recognizable blue-gray color of the Russian Blue is actually a result of a genetic mutation that affects the production of the pigment melanin. Overall, the Russian Blue is a beautiful and striking breed with a distinctive coat color that makes it a popular choice among cat enthusiasts.

Figure 8: Step 4 - final response production. We present the prompt used for short-text generation.



[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider correctness and helpfulness. You will be given a reference answer and the assistant's answer. Begin your evaluation by comparing the assistant's answer with the reference answer. Identify and correct any mistakes. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]
{question}

[The Start of Reference Answer]
{reference}
[The End of Reference Answer]

[The Start of Assistant's Answer]
{prediction}
[The End of Assistant's Answer]

Figure 9: The prompt fed to GPT4 for evaluating model predicted rationales on GSM8K.