Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework

Anonymous ACL submission

Abstract

As large language models (LLMs) have become the norm in NLP, demonstrating good performance in generation and reasoning tasks, one of its most fatal disadvantages is the lack of factual correctness. Generating unfactual texts not only leads to lower performances but also degrades the trust and validity of their applications. Chain-of-Thought (CoT) prompting improves trust and model performance on complex reasoning tasks by generating interpretable reasoning chains, but still suffers from factuality concerns in knowledge-intensive tasks. In this paper, we propose the Verify-and-Edit framework for CoT prompting, which seeks to increase prediction factuality by post-editing reasoning chains according to external knowledge. Building on top of GPT-3, our framework lead to accuracy improvements in multiple open-domain question-answering tasks.[1]

1 Introduction

Large Language Models (LLMs) have become the new norm in many downstream NLP tasks. In utilizing these LLMs, Chain-of-Thought (CoT) prompting (Wei et al., 2022) is found to improve performances for tasks that require complex reasoning, such as math word problems, commonsense reasoning, and symbolic manipulation. At the same time, it is able to generate interpretable reasoning chains. Recent work further explored how to use these reasoning chains to select better predictions. However, the primary focus of these methods has been to improve end-task performance by utilizing generated CoTs as-is. For example, Ye and Durrett (2022) train a calibrator that tunes prediction probabilities based on rationale scores; Wang et al. (2022) sample multiple reasoning paths to find the most common (consistent) prediction. Only a few, such as Creswell et al. (2022) and Zhou et al. (2022), have explored ways to improve the quality of CoTs themselves.
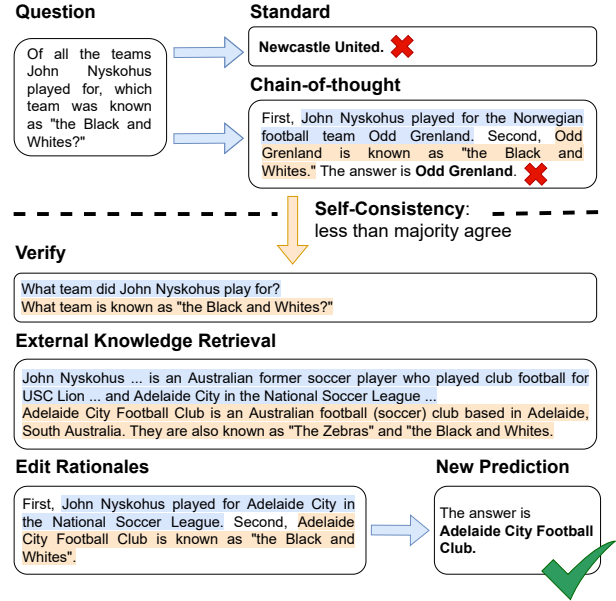

Figure 1: The Verify-and-Edit framework consists of five steps: (1) pass predictions with lower-than-average consistency to the next stages while leaving highly consistent predictions as-is; (2) produce verifying questions; (3) retrieve external knowledge; (4) edit rationales with informed answers; and (5) produce new predictions.

In fact, improving the CoT quality could be beneficial in enhancing both interpretability and end-task performance. Ye and Durrett (2022) point out that explanations judged as good by humans often indicate more accurate predictions. Intuitively, a better set of CoT prompts could provide better grounding and logically consistent thought processes, thus leading to more accurate predictions.

To improve generation quality, one important aspect is *factual correctness*, which is currently one of the most fatal drawbacks of LLMs (OpenAI-Blog, 2022). In answering user queries, LLMs such as GPT-3 (Brown et al., 2020) tend to make up facts and details, which is now flagged as a primary warning in their API usage. As a major use case of LLMs is the prospect of replacing traditional

---

[1] For reproducing the results, our codebase is available here

search engines and usage for more direct information access through question-answering, factuality concerns could largely undermine their validity and degrade users' level of trust (Marcus, 2022). Fixing this issue is challenging and the concerns still persist even after the models are instruction-tuned with human feedback (Ouyang et al., 2022). This is because the source of truth can be unavailable during the finetuning process (OpenAI-Blog, 2022).

Thus, it is of urgent concern to better control the generation and increase the factual correctness of predictions. As LLMs could fail to recall accurate details when functioning as a knowledge base (Ye and Durrett, 2022; Creswell et al., 2022), if possible, knowledge from external sources could be introduced as assistance. Assisted thought process is also common in human reasoning: when humans answer questions, they often search (or revisit) external knowledge sources for supporting facts in order to refresh their (internal) memory.

Inspired by this, in this work we propose a **Verify-and-Edit** (VE) framework to post-edit the reasoning chains for more factually aligned predictions. As shown in Fig. 1, we first select uncertain instances to edit, which have a less-than-majority-agree consistency. These instances, as implied in Wang et al. (2022), often consist of plausible-sounding statements, such as the sentence "John Nyskohus played for the Norweigian football team Odd Greenland" in Fig. 1. When editing, we first generate a question to verify this detail, such as "What team did John Nyskohus play for?" Then, to answer this query, we introduce external knowledge through open-domain retrieval systems. For example, the fact "John Nyskohus ... played for Adelaide City.." is retrieved in this instance. Then, the rationales are edited by providing the retrieved facts in the prompts as memory refreshments. Thus, the edited rationales could be updated corresponding to the retrieved facts (Fig. 1). Given the edited rationales, the new prediction is generated, which considers more factually aligned reasoning traces.

To our knowledge, our work is the first to post-edit CoT-style reasoning chains to enhance prediction performance. We perform experiments on two open-domain Question Answering (QA) tasks that require reasoning: Adversarial HotpotQA (Yang et al., 2018) and 2WikiMultihop (Ho et al., 2020). We also test its performance on the Fact Verification task using Fever (Thorne et al., 2018). We find that the model is able to benefit from more factual rea-

soning chains, thus generating more accurate predictions. For example, for open-domain QA, our model demonstrates 3.8x accuracy improvement compared to similar retrieval-augmented models on AdvHotpot. On 2WikiMultihop, Verify-and-Edit reaches 33.6% accuracy with open-domain search, while CoT Self-Consistency stands at 27.7%.

2 Related Work

Chain-of-Thought or CoT (Wei et al., 2022) is a prompting method for improving the reasoning abilities of LLMs, which enables LLMs to decompose complex problems into multiple intermediate steps. CoT provides interpretability and has been proven to be more capable of solving complex problems than standard prompting methods.

However, hallucination is a long-standing problem in NLP, especially for LLMs, which has drawn significant attention from the research communities. The decoding process of LLMs is auto-regressive, which unavoidably makes it output nonfactual content without controlled generation (Ye and Durrett, 2022; Wiegreffe et al., 2022). As such, the lack of supporting facts during the generation process of CoT could largely undermine the validity of the final answer (Golovneva et al., 2022). Ye and Durrett (2022) demonstrate that the accuracy of the final answers largely correlates with the factuality and consistency of the reasoning explanations. The commonly proposed methods to improve the factuality of CoT reasoning process can be grouped into two categories: prompt engineering and result calibration.

Prompt engineering methods are usually applied to guide LLMs to generate better intermediate reasoning explanations. *ReAct* (Yao et al., 2022), which is the most comparable to our work, synergizes reasoning and acting in LLMs, where reasoning steps help the model induce and update actions, while action steps allow the model to consult additional information from Wikipedia for a factuality check. Compared to *ReAct*, we generate more natural and conversational CoTs for better interpretability and easier learning. As such, our framework requires a much shorter prompt to learn. Press et al. (2022) propose *self-ask* by instructing the LLM to explicitly ask itself (and then answer) follow-up questions before answering the initial question. One natural way of solving a complex problem is to decompose the problem into sub-problems and solve them sequentially. Zhou et al.

(2022) adopt the idea and propose *least-to-most* prompting. However, both *self-ask* and *least-to-most* prompting still rely on repetitively retrieving internal knowledge learned by the LLM instead of connecting to external knowledge. Thus, their ability to improve factuality is limited.

Result calibration functions on the output of the LLMs. Ye and Durrett (2022) train a calibrator to calibrate the weights of the final answers based on the factuality and consistency of the generated explanations, which efficiently improves the results. The decoding method in CoT is naive greedy, which simply outputs the next token with the highest probability. Wang et al. (2022) propose a *self-consistency* decoding method, which samples a diverse set of reasoning paths and then selects the most consistent answer by marginalizing out the sampled reasoning paths. *Selection-Inference (SI)* (Creswell et al., 2022) framework is another state-of-the-art method that exploits LLMs as general processing modules. Out of all the methods, it is also the first to systematically improve the factual correctness of CoTs in order to predict more accurately. It alternates between selection and inference to generate a series of interpretable, causal reasoning steps leading to the final answer, which is proven to be efficient. However, it is not designed for open-domain or commonsense question answering.

Moreover, another comparable line of work has been exploring retrieval-augmented language model pretraining (REALM) (Guu et al., 2020), which first retrieves documents from an external knowledge source and then utilizes retrieved documents to process question-answering tasks. Lazaridou et al. (2022) propose to include Google search results of the question in the prompt to improve the factuality of the generated answer. However, such methods may fail in complex questions as it does not utilize the reasoning capability of LLMs. Thus, we consider retrieval-augmented reasoning paths as a natural way to increase factual alignment.

3 Verify-and-Edit Framework

Our goal is to make LLMs generate more factual reasoning chains with CoT prompting assisted with external knowledge, thereby also improving prediction accuracy of the final answer. We hypothesize that this can enhance LLMs' capability to solve complex knowledge-intensive tasks that require multiple reasoning steps to arrive at an answer.

Generally, we hope to follow the human reasoning process: when a person answers a question, if he/she is unsure, he/she would search for a supporting fact and consider it before giving the final answer. Thus, we could separate the Verify-and-Edit (VE) framework into 3 different stages: finding uncertain predictions, editing their rationales by searching for supporting facts, and using the edited rationales to generate final answers (Fig. 1). In designing the stages, we hope to maximally preserve the LLMs' biggest advantage: their open-generation and reasoning ability. And we aim to design tasks and setups as natural and conversational as possible, thus making it easy to understand for humans and LLMs which are trained with natural texts.

3.1 Deciding when to edit

How can we identify when a model is unsure of its prediction? The self-consistency method (Wang et al., 2022) provides a solution. In sampling diverse reasoning paths and answers, self-consistency is found to be highly correlated with accuracy, suggesting that it could provide an uncertainty estimate and confer abilities for the model to "know when it doesn't know". Thus, we begin the VE framework by using the consistency method to sample n diverse reasoning paths for a prediction task. The highly consistent predictions are left as-is. When consistency is lower than $\lceil n/2 \rceil$, *i.e.* the majority cannot agree on the same answer, we label it as "uncertain".

3.2 How to edit a specific rationale

The rationale, *i.e.* the thought process (CoT), could be viewed in two parts: facts and reasoning which combines facts to derive a new claim. Thus, we consider improving the CoT from both aspects.

• **Facts** To make the thought process more factually correct, we search for supporting facts in external knowledge sources (*e.g.* Wikipedia, Google).

First, to mimic a human's query when searching for validating facts, a natural question is generated to verify the rationale. For this, we use the in-context learning capability of the same LLM. The original question and the rationale are both provided in the prompt for verifying question generation to ensure that it asks for the most relevant information required to answer the original question, instead of other entities in the rationale. For example, if the rationale (wrong) is "the US pres-

Algorithm 1 Verify-and-Edit

Require: The original question q; An n-shot CoT prompt p_{cot}
Require: An LLM $f(\cdot)$; LM number of completions n; LM decoding temperature τ
Require: An external knowledge retrieval model $g(\cdot)$
Require: n-shot prompts for verifying question generation (p_{vq}) and answer generation (p_{va})

    $R, A \leftarrow f(p_{cot}, q, n, \tau)$ $\triangleright$ Generate a set of reasonings (R) and answers (A).
    $s_{sc}^* \leftarrow \max P(a|p_{cot}, q), a \in A$ $\triangleright$ The highest self-consistency score among all answers.
    $r^*, a^* \leftarrow \arg\max P(a|p_{cot}, q), a \in A$ $\triangleright$ Reasoning and answer with highest self-consistency.
    if $s_{sc}^* < \lceil \frac{n}{2} \rceil$ **then** $\triangleright$ Edit reasoning with a less-than-majority-agree consistency.
        for $o_i \in r^*$ **do** $\triangleright$ Edit each sentence in the reasoning.
            $u \leftarrow f(p_{vq}, q, o_i)$ $\triangleright$ Generate verifying question.
            $v \leftarrow g(u)$ $\triangleright$ Retrieve external knowledge.
            $w \leftarrow f(p_{va}, u, v)$ $\triangleright$ Generate verifying answer.
            $o_i \leftarrow w$ $\triangleright$ Edit original reasoning sentence with verifying answer.
        end for
        $a^* \leftarrow f(p_{cot}, q, r^*)$ $\triangleright$ Generate final answer with edited reasoning.
        return a^*
    else if $s_{sc}^* \geq \lceil \frac{n}{2} \rceil$ **then** $\triangleright$ Answer with high consistency is left as-is.
        return a^*
    end if

ident born on 4 August 1961 is John Kennedy." and the original question is "who is the spouse of the US president born on 4 August 1961", we expect the generated verifying question to be: "Who is the US president born on 4 August 1961?" instead of "When is John Kennedy's birthday?" By generating a relevant question instead of directly querying with the generated rationale, we eliminate potential noise brought by incorrect fact generation. In the example above, if one retrieves using the wrong claim "the US president born on 4 August 1961 is John Kennedy", the incorrect entity "John Kennedy" may obfusticate the search process.

In this paper, we use relevant contexts retrieved from 3 systems: (*i*) DrQA (Chen et al., 2017), an open-domain question-answering system; (*ii*) Wikipedia search of relevant pages; and (*iii*) Google search, which demonstrates possibilities of combining LLMs and search engines.

As the retrieved contexts from a retrieval system could be longer than desired, we use a pre-trained LM to rank and select the top-k sentences most similar to the verifying question query.

● **Reasoning** While methods such as Selection-Inference (Creswell et al., 2022) directly use retrieved facts as rationales, they are usually too verbose, longer than desired, or contain irrelevant details. Ye and Durrett (2022) have made similar observations: directly using supporting sentences

is usually too verbose and not sufficient.

To obtain more relevant and logical rationales, we again utilize a natural and generative approach, as reasoning abilities are believed to be already built into LLMs (Wei et al., 2022). In particular, by feeding in prompts in the format of "question, rationale, answer", the LLM learns to reason for a few steps before answer generation. Upon investigating the original rationales, we observe that, even when they contain incorrect facts, the logical reasoning component seems to be generally intact. Thus, we use the verifying questions (as logic) and retrieved facts (as information) to generate informed answers. The informed answers are then composed into a new rationale, providing potentially a more factual CoT.

3.3 Answering again

Finally, with the post-edited CoT, new answers are generated by prompting the LLM. A pseudocode of the overall procedure is given in Alg. 1, and illustrated with an example in Fig. 1 . We can see that, by allowing the LLM to incorporate external knowledge, our method could result in more factually-grounded rationales. When prompted into the LLM as a CoT, it could bring in the information necessary to make a new prediction, which was originally not remembered correctly by the model.

Compared to specifically designed prompts such as ReAct (Yao et al., 2022), the Verify-and-Edit

framework is simple and arguably more natural. Its conversational nature could allow humans to better understand the model's thought processes and have the potential for users to naturally interfere and revise at any stage of inference. In the experiments presented next, we also observe that such a setup is effective in mitigating factuality concerns and boosting end-task performances.

4 Experiment Setup

4.1 Reasoning tasks

As the Verify-and-Edit framework offers more knowledge-grounded reasoning steps, it should benefit tasks that fulfill the following two properties: (*i*) reliant on multi-hop reasoning to arrive at a later prediction, thus depending on rationale generation, and (*ii*) open-domain, thus needing to interact with an external knowledge source.

Therefore, we validate the approach on three datasets: (*i*) **Adversarial HotpotQA** (Yang et al., 2018), a multi-hop question answering dataset. We use the challenging subset proposed by Ye and Durrett (2022), where the correct and incorrect predictions are balanced using their model. (*ii*) **2Wiki-Multihop** (Ho et al., 2020) a multi-hop question-answering dataset exploiting the structured format in Wikidata and use logical rules.[2] (*iii*) **Fever** (Thorne et al., 2018), a fact verification dataset that labels claims as "SUPPORTS", "REFUTES", or "NOT ENOUGH INFO" based on evidence paragraphs from Wikipedia. Similar to the HotpotQA setup, we sample a challenging set by balancing the samples where GPT3 CoT makes correct and incorrect predictions. Details on the processing and use of the datasets can be found in Appendix A.

4.2 Compared methods

To provide the most state-of-art performance estimates, we utilize the GPT-3 instruct series API `text-davinci-003` (Ouyang et al., 2022), the strongest and most up-to-date model at the time of experiments, as a backbone. The cost of experiments is stated in Appendix B.

Adversarial HotpotQA and 2WikiMultihop experiments used 6-shot and Fever used 3-shot in-context learning, as Fever questions are shorter and easier to learn. We use the manual annotations provided for HotpotQA by Ye and Durrett (2022) and manually annotate few-shot examples

for 2WikiMultihop and Fever in a similar format. Full prompts for baseline and our methods are provided in Appendix C.

Baselines   To provide a more comprehensive overview of where our framework stands, we use the following baselines:

1. **Standard Prediction** (Standard): Directly predicting the label based on input, given the same number of in-context learning examples.
2. **Original CoT** (Wei et al., 2022): Predicting the label after generating the explanation.
3. **CoT with Self-Consistency** (CoT-SC) (Wang et al., 2022): Sampling 5 CoT trajectories with a decoding temperature of 0.7, which is recommended by the paper.
4. **Calibrator** (Calib.) (Ye and Durrett, 2022): A calibrator that tunes the probabilities of a prediction based on the score of its prediction.
5. **ReAct** (Yao et al., 2022): A reason-and-act framework that utilizes an external Wikipedia API. For this baseline, we use the reported results in the original paper, which uses the PaLM model (Chowdhery et al., 2022), whose performance is similar to GPT-3.[3] To add a more justified perspective, we report its performance improvement gained on top of the CoT-SC baseline. [4]

Verify-and-Edit (VE)   In implementing the VE framework, the same consistency baseline is employed to estimate when the model is uncertain. As stated in §3.1, we edit all instances with a self-consistency score below $\lceil n/2 \rceil$, where n is the number of sampled paths. Then, the verifying questions are produced using a 2-shot[5] setup with in-context learning. The verifying answers are produced using the same number of examples in original answer generation and greedy decoding.

To study the effect of knowledge retrieval systems on the results, we use four systems:

1. **Wikipedia-API** (wiki): Searching for the query entities and selecting top sentences from their Wikipedia pages.
2. **DrQA** (Chen et al., 2017): A pre-trained open-

---

[2]We randomly sample 1,000 samples out of 12,576 dev samples for cost considerations.

[3]We could not use PaLM as it is not open-sourced.

[4]it is worth noting that ReAct conducted experiments on the entire dataset, where we used a sampled version (see §4.1).

[5]As we observe that question generation quality does not vary too much as in-context examples increase, we select the shortest prompt that is able to generate reasonable questions to reduce cost.
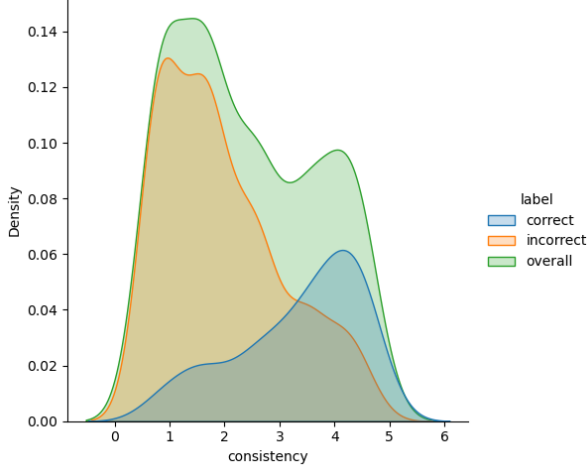
Figure 2: Kernal density estimation plots for consistency on the Adversarial **HotpotQA** dataset. With kernal estimation, the curve extends its true distribution's range, which is from 0 to 5 (as we sampled 5 paths).

Method	knowledge	EM	ΔEM	AUC
CoT-SC → ReAct	Wiki.	34.2%	+0.8%	-
ReAct → CoT-SC	Wiki.	35.1%	<u>+1.7%</u>	-
Standard	-	23.1%	-	43.24
CoT	-	31.8%	-	38.30
CoT-SC	-	31.2%	-	34.97
CoT-SC + Calib.	Dataset	-	-	<u>49.00</u>
CoT-SC + VE	Wiki.	35.7%	+4.5%	45.62
CoT-SC + VE	DRQA	36.0%	+4.8%	46.06
CoT-SC + VE	Google	<u>37.7%</u>	<u>+6.5%</u>	47.98
CoT-SC + VE	Dataset	**56.8%**	**+25.6%**	**60.94**

Table 1: Results on the Adversarial **HotpotQA** dataset. The best result for each model is underlined and the best result overall is bolded. ΔEM represents the improvement on Exact Match from the CoT-SC baseline. The top two rows uses the PaLM model and the rest uses the GPT-3 davinci-003 model.

domain QA model that combines bigram hashing, TF-IDF matching, and a multi-layer recurrent neural network model. We only utilize the contexts retrieved from it.[6]

3. **Google**: Using top-k search results produced by Google as assistive contexts. This result is interesting in providing possibilities in combining search engines and LLMs.

4. **Dataset**: Selecting from the set of paragraphs provided in Adversarial HotpotQA and 2Wiki-MultihopQA, which includes ground-truth supporting contexts and distractor paragraphs. This is similar to an oracle setup, which provides an upper bound of the performance boost, assuming we have a good retrieval system.

For 1, 2, and 4, after retrieving, we select the top 3 sentences most similar to the query ranked by the pre-trained Sentence BERT model (Reimers and Gurevych, 2019) as context.

5 Results and Analysis

5.1 Using Self-Consistency: know when it doesn't know

For the first step in the Verify-and-Edit framework, consistency is used to measure the model's confidence in a prediction. Aligned with the findings from Wang et al. (2022), we hypothesize that when

the consistency is low, the model is more uncertain and thus more likely to generate inaccurate predictions. To test whether this hypothesis holds, we plot the kernal density estimation plots for consistency distribution on the Adversarial HotpotQA dataset. As shown in Fig. 2, the incorrect samples show a left-skewed consistency distribution, where most incorrect predictions have low consistencies. On the other hand, the distribution of correct predictions shows a right-skewed tendency, where there are very few incorrect samples with higher consistencies. This effectively validates our hypothesis.

In the main experiments, we use $\lceil n/2 \rceil$ as a majority threshold and edit all samples below it, which is at 3. To show the effects of different thresholds on the framework's performance, we also provide an ablation study later.

5.2 Results on HotpotQA

Reported in Table 1, we observe that CoT improves on top of the Standard few-shot setting. CoT-SC, on the other hand, does not demonstrate a good improvement on the baseline. Using the calibrator from Ye and Durrett (2022), AUC is improved as it learns to calibrate the answer weights based on ground-truth contexts provided in the dataset. Thus, it should be compared with the last setup of VE, where we use dataset knowledge. In comparison, the calibrator results in a lower AUC and cannot improve the accuracy as it does not generate alternative answers in open-domain settings.

Using the Verify-and-Edit framework, the retrieval systems Wikipedia and DrQA could generate an improvement of 4.5% and 4.8% respectively on top of the baseline, which is 2x the highest EM

---

[6]We selected DrQA by first conducting small-scale experiments with different open-domain QA models, including DPR (Karpukhin et al., 2020). DrQA is found to yield better performance. Thus, we consistently use it.

Method	knowledge	EM	ΔEM	AUC
Standard	-	16.9%	-	35.89
CoT	-	28.4%	-	16.64
CoT-SC	-	27.7%	-	17.16
CoT-SC + Calib.	Dataset	-	-	24.13
CoT-SC + VE	Wiki.	33.1%	+5.4%	28.32
CoT-SC + VE	DRQA	31.1%	+3.4%	27.75
CoT-SC + VE	Google	33.6%	+5.9%	30.06
CoT-SC + VE	Dataset	**37.2%**	**+9.5%**	32.28

Table 2: Results on **2WikiMultiHopQA** dataset. ΔEM represents the improvement on Exact Match from the CoT-SC baseline. All experiment uses the GPT-3 davinci-003 model.

Method	knowledge	Accuracy	Δ Accuracy
CoT-SC → ReAct	Wiki.	-	+4.2%
ReAct → CoT-SC	Wiki.	-	+1.6%
Standard	-	46.8%	-
CoT	-	50.0%	-
CoT-SC	-	52.0%	-
CoT-SC + Calib.	-	33.7%	
CoT-SC + VE	Wiki.	53.6%	+1.6%
CoT-SC + VE	DRQA	53.3%	+1.3%
CoT-SC + VE	Google	53.9%	+1.9%

Table 3: Results on **Fever** dataset. ΔAccuracy represents the improvement on Accuracy from the CoT-SC baseline. The top two rows uses the PaLM model and the rest uses the GPT-3 davinci-003 model.

improvement for ReAct (1.7%). When we combine the search engine results from Google into the framework, the EM is increased by 6.5%, which is 3.8x the ReAct result. This shows a promising method for combining search engines and LLMs, which is a popular direction now. Search engines return factual results, but are less powerful in queries that require reasoning. On the other hand, LLMs are powerful in reasoning and abstraction but tend to generate plausible-sounding but incorrect statements (OpenAI-Blog, 2022). To combine the best of both worlds, we could utilize the long memory of LLMs, as many users have reported that GPT is able to remember inputs mentioned earlier in the dialogue. By providing factual results from the search engines as a memory refreshment, GPT is able to generate better and more factual predictions.

Then, when we use the adversarially augmented paragraphs provided in the dataset, the model is able to demonstrate very high EM (56.8%) and AUC (60.94) at the same time. This setup shows that, if we have a highly compressed set of contexts and a nearly-ideal retrieval system, the Verify-and-Edit framework could potentially result in very strong performances.

5.3 Results on 2WikiMultiHop

As shown in Table 2, our method demonstrates even stronger performances on 2WikiMultiHop compared to HotpotQA. The Verify-and-Edit framework with open-domain retrieval is able to generate a high accuracy improvement, ranging from 3.4% to 5.9%. Selecting from paragraphs provided in the dataset, which includes supporting evidences and irrelevant paragraphs, the accuracy improvement is further increased to 9.5%. The calibrator, on the other hand, uses the dataset provided paragraphs but still lags behind all variations of our

Verify-and-Edit framework.

5.4 Results on fact verification

Results on the Fever dataset are shown in Table 3. As the reasoning required by the Fever dataset is less multi-hop compared to HotpotQA and 2WikiMultiHop, we anticipate that it should demonstrate lower improvements compared to the other two.

In the Fever dataset, the calibrator method completely fails, decreasing to 33.7%: it calibrates the prediction scores based on factuality estimates, which is produced by examining the overlap between the reasoning path and the provided context. However, in such Fact Verification datasets, there is no provided contexts. Thus, we calibrate using the original claim, which results in bad performances. It shows here that one limitation of the calibrator method is that it only applies to cases with provided relevant contexts.

Even though this task does not require much reasoning, employing the Verify-and-Edit framework, we are able to observe consistent improvements over the baseline method. Similar to before, the Wikipedia retrieval is able to result in a larger improvement over DrQA, and Google search improves further at 1.9%.

Compared to our method, ReAct is able to demonstrate a larger improvement on Fever. First of all, it has been mentioned before that Fever is less suited for the Verify-and-Edit framework as it requires less reasoning to solve the task. Secondly, ReAct prompts are much longer than our prompts, requiring more computational costs.

5.5 Evaluating the reasoning chains with human study

To closely examine the faithfulness of the generated reasoning chains, we also conduct a small-

# Examples	Cohen κ	CoT-SC	Ours	Tie
50	0.25	17%	**53%**	30%

Table 4: Human study for factuality of CoTs on the HotpotQA dataset. "Ours" refers to the Verify-and-Edit model with Google retrieval.
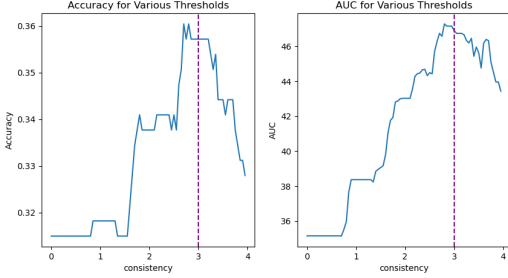

Figure 3: Ablation study on the effect of various consistency thresholds on task performances on Adversarial HotpotQA

scale human study experiment. During the experiment, two human volunteers are shown 50 randomly selected questions with generated reasoning chains from CoT-SC and Verify-and-Edit on the HotpotQA dataset. They are then asked to select the more factually consistent one. Volunteers are encouraged to use search engines as assistance. A detailed description on the setup is described in Appendix D.

Shown in Table 4, humans select the reasoning chains produced by Verify-and-Edit as more factually consistent 53% of the time, compared to 17% for the CoT-SC baseline. The Cohen κ is at 0.25, showing fair agreement between the two annotators (McHugh, 2012). The annotators used Google search as an assistive tool 100% of the time, which shows the necessity of introducing external knowledge.

Moreover, human annotations in this case require a lot of efforts. Annotators report 1.5 minutes on average to validate one data point. Thus, automating the Verify-and-Edit process is of benefits as an assistive tool to reduce human labor.

To observe the qualitative effects of the Verify-and-Edit framework in detail, we also include several interesting examples in Appendix E, which show the effectiveness of our framework in correcting the original claims.

5.6 Ablation study: editing at different consistency thresholds

In the Verify-and-Edit framework, the only hyper-parameter to select is the consistency threshold. Similar thresholds also exists in ReAct (Yao et al., 2022), where the CoT $\rightarrow$ ReAct method is to employ ReAct-style prompting when "the majority answer among n CoT-SC samples occurs less than n/2 times". Using majority counts, however, is less fine-grained compared to using the original consistency formulated with log probablities. Thus, we employ the original score proposed by Wang et al. (2022), which is the unnormalized answer probabilities marginalized over the rationales' log probabilities. To mimic a majority-vote threshold, we select $\lceil n/2 \rceil$, where n is the number of sampled paths.

To study the effect of adjusting the consistency threshold on our framework, we show the ablation results of Adversarial HotpotQA in Fig. 3. As the threshold increases, accuracy first increases, reaching a peak close to $\lceil n/2 \rceil$, which is 3, before decreasing. The AUC scores demonstrate a similar trend.

As shown in Fig. 2, when consistency is larger than majority ($\lceil n/2 \rceil$), there are usually more correct predictions rather than incorrect predictions, and vice versa. Thus, as we increase the consistency threshold from 0 to $\lceil n/2 \rceil$, more uncertain and possibly incorrect samples are getting edited by introducing external knowledge. As we go beyond the ideal threshold $\lceil n/2 \rceil$, we are mostly re-editing correct samples, and the introduced noise may disrupt the original reasoning chains.

Thus, we recommend a consistency threshold at $\lceil n/2 \rceil$ as an ideal level.

6 Conclusions

In this paper, we introduce a Verify-and-Edit framework for open-domain question-answering. It is a first attempt to post-edit CoT-style reasoning chains for better end-task performance. By combining knowledge retrieval with reasoning, the framework edits CoTs in a natural and conversational way, which enhances prediction factuality. Combined with Google search, the framework also shows a promising direction that combines the open-generation ability of state-of-art LLMs with the updated facts provided by search engines.

Limitations

There are a few limitations to the current framework. Firstly, Verify-and-Edit works the best for open-domain question-answering tasks that require complex reasoning. Less complex datasets or commonsense datasets that do not require knowledge retrieval may not result in high improvements. Secondly, it is most ideal to edit a group of mostly incorrect samples, which we try to select by using consistency. Thus, our method is reliant on the consistency method's performance and its abilities to separate correct and incorrect predictions. Most often, it can demonstrate a larger improvement with a more challenging set of examples.

To address these limitations, we plan to work on reducing the noise brought in the rationale-editing stage with reinforcement learning as a follow-up.

Ethics Statement

The Verify-and-Edit framework can mitigate potential ethical concerns of LLM generation surrounding hallucinations and unfactual details. Some persisting concerns include: (1) As the framework uses google as one of the retrieval methods, it could retrieve potentially toxic information that exists in google search results. (2) As the framework uses GPT3 as a backbone, it could suffer from existing ethical concerns of GPT3, such as responding to toxic queries or exhibiting biased behavior.

For knowledge retrieval, we used Wikipedia corpus and google search results. Permission is granted to copy, distribute and/or modify Wikipedia's text under the terms of the Creative Commons Attribution-ShareAlike 3.0 Unported License. For google search results, scraping publicly accessible data is legal considered by the U.S. appeals court.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering.

Gary Marcus. 2022. Is chatgpt really a "code red" for google search?

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

OpenAI-Blog. 2022. Chatgpt: Optimizing language models for dialogue.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Appendix for "Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework"

A   Dataset Processing

A.1   Adversarial HotpotQA

The Adversarial HotpotQA subset is formed in Ye and Durrett (2022), who processed the original set in a few ways: (1) Context length is reduced to make it better fit the purpose of testing in-context learning. (2) Set of adversarial contexts is reduced to two ground truth supporting paragraphs and two adversarial paragraphs, instead of using all eight distractors. Each paragraph is further simplified by only keeping relevant sentences needed for answering the question (or distracting the prediction) (3) A challenging test set of 250 examples is formed by balancing the mix of examples on which prompted `text-davinci-001` (which is used at their time of experiments) to make correct and incorrect predictions. This is done by first running few-shot inference over 1000 examples, and then randomly sampling 125 examples with correct and incorrect predictions, respectively. The subsampled dataset is available publicly at the github for Ye and Durrett (2022).

The HotpotQA dataset is distributed under the CC BY-SA 4.0 license, which allows for modification and research use.

A.2   2WikiMultihopQA

For cost concerns, we randomly subsample 1,000 out of the dev set of 12,576 samples, which provides a reasonable estimate. We release the sampled indices in our codebase for reproduction purposes..

The 2wikimultihop dataset is licensed under the Apache License 2.0, which allows for modification and research use.

A.3   Fever

To mimic the Adversarial HotpotQA setup, we run the CoT baseline for 3,000 samples and randomly sample 1,000 by balancing the number of right and wrong predictions. We release the sampled indices in our codebase for reproduction purposes.

Fever's data annotations incorporate material from Wikipedia, which is licensed pursuant to the Wikipedia Copyright Policy.

B   Experiment Costs

For the experiments, we use the API for `text-davinci-003`. The costs for inferencing the LLM is $0.02/1K tokens. We spent in total 273$.

C   Prompts Used

C.1   HotpotQA

C.1.1   Few-shot prompt

Q: This British racing driver came in third at the 2014 Bahrain GP2 Series round and was born in what year
A: 1991

Q: What band did Antony King work with that formed in 1985 in Manchester?
A: Simply Red

Q: How many inhabitants were in the city close to where Alberta Ferretti's studios was located?
A: 146,606

Q: TLC: Tables, Ladders & Chairs was a wrestling event featuring which American wrestler and rapper in the main event?
A: John Felix Anthony Cena

Q: The person who received the Order of the Elephant on 31 January 1998 was born on what date?
A: 27 April 1967

Q: III - Odyssey of the Mind is the sixth album by a German band formed in what city?
A: Düsseldorf

Q: [Question]
A:

C.1.2   CoT, CoT-SC prompt

Q: This British racing driver came in third at the 2014 Bahrain GP2 Series round and was born in what year
A: First, at the 2014 Bahrain GP2 Series round, DAMS driver Jolyon Palmer came in third. Second, Jolyon Palmer (born 20 January 1991) is a British racing driver. The answer is 1991.

Q: What band did Antony King work with that formed in 1985 in Manchester?
A: First, Antony King worked as house engineer for Simply Red. Second, Simply Red formed in 1985 in Manchester. The answer is Simply Red.

11

Q: How many inhabitants were in the city close to where Alberta Ferretti's studios was located?
A: First, Alberta Ferretti's studio is near Rimini. Second, Rimini is a city of 146,606 inhabitants. The answer is 146,606.

Q: TLC: Tables, Ladders & Chairs was a wrestling event featuring which American wrestler and rapper in the main event?
A: First, TLC: Tables, Ladders & Chairs was a wrestling event featuring John Cena in the main event. Second, John Cena is an American wrestler and rapper. The answer is John Felix Anthony Cena.

Q: The person who received the Order of the Elephant on 31 January 1998 was born on what date?
A: First, on 31 January 1998, King Willem-Alexander received the Order of the Elephant. Second, Willem-Alexander was born on 27 April 1967. The answer is 27 April 1967.

Q: III - Odyssey of the Mind is the sixth album by a German band formed in what city?
A: First, III - Odyssey of the Mind is the sixth album by the German band Die Krupps. Second, Die Krupps is formed in Düsseldorf. The answer is Düsseldorf.

Q: [Question]
A:

C.1.3 Verifying Question Generation prompt

Write a question that asks about the answer to the overall question.

Overall Question: The Sentinelese language is the language of people of one of which Islands in the Bay of Bengal?
Answer: The language of the people of North Sentinel Island is Sentinelese.
Question: What peopleś language is Sentinelese?

Overall Question: Two positions were filled in The Voice of Ireland b which British-Irish girl group based in London, England?
Answer: Little Mix is based in London, England.
Question: What girl group is based in London, England?

Overall Question: [original question]
Answer: [rationale sentence to edit]
Question:

C.1.4 Verifying Answer Generation (Rationale Editing) prompt

Barnes House (born 20 January 1969) is a British racing driver, currently driving for Renault Sport F1 Team in the Formula One World Championship. Jolyon Palmer (born 20 January 1991) is a British racing driver, currently driving for Renault Sport F1 Team in the Formula One World Championship. Ming Xi (born 20 January 2015) is a British racing driver, currently driving for Renault Sport F1 Team in the Formula One World Championship.
The 2014 Bahrain GP2 Series round was a pair of motor races held on 6 and 7 April 2014 at the Bahrain International Circuit in Sakhir, Bahrain as part of the GP2 Series. Julián Leal finished second for the Carlin team and DAMS driver Jolyon Palmer came in third.
Q: This British racing driver came in third at the 2014 Bahrain GP2 Series round and was born in what year
A: This British racing driver came in third at the 2014 Bahrain GP2 Series round and was born in 1991..

Antony King (born 1974) is a British live audio engineer for Depeche Mode and Nine Inch Nails. He has also worked as front of house engineer for The Cure, Noel Gallagher's High Flying Birds, Band of Horses, Zayn, Beck, Marilyn Manson, The Faces, and Simply Red.
Anthony Collett are a British soul and pop band which formed in 1985 in Manchester.
Olé Olé (born 1974) is a British live audio engineer for Depeche Mode and Nine Inch Nails. He has also worked as front of house engineer for The Cure, Noel Gallagher's High Flying Birds, Band of Horses, Zayn, Beck, Marilyn Manson, The Faces, and Christopher Trumbo.
Simply Red are a British soul and pop band which formed in 1985 in Manchester.
Q: What band did Antony King work with that formed in 1985 in Manchester?
A: Antony King work with the band Simply Red, which was formed in 1985 in Manchester..

Alberta Ferretti (Cattolica, 1950) is an Italian fashion designer and dressmaker. Her showroom is in Milan, Italy but her studio is in the village of Cattolica, near Rimini, Italy.
Rimini (] ; Romagnol dialect: "Rémin"; Latin: "Ariminum") is a city of 146,606 inhabitants in

the Emilia-Romagna region of northern Italy and capital city of the Province of Rimini.

Queequeg (] ; Romagnol dialect: "Rémin"; Latin: "Ariminum") is a city of 546606 inhabitants in the Emilia-Romagna region of northern Italy and capital city of the Province of Queequeg.

Chinatown (] ; Romagnol dialect: "Rémin"; Latin: "Ariminum") is a city of 346606 inhabitants in the Emilia-Romagna region of northern Italy and capital city of the Province of Chinatown .

Q: How many inhabitants were in the city close to where Alberta Ferretti's studios was located?

A: 146,606 inhabitants were in the city close to where Alberta Ferretti's studios was located..

[contexts]

Q: [verifying question]

A:

C.2 2WikiMultihop

C.2.1 Few-shot prompt

Q: Which film was released earlier, Kistimaat or I'M Taraneh, 15?

A: I'M Taraneh, 15

Q: What is the date of death of the composer of film Baalaraajana Kathe?

A: 27 May 1980

Q: Who is the spouse of the director of film Alive (1993 Film)?

A: Kathleen Kennedy

Q: Who lived longer, Edward Frederick Sanderson or Forrest Towns?

A: Edward Frederick Sanderson

Q: Which country the director of film Battle Circus (Film) is from?

A: American

Q: Which country the director of film In The Fog is from?

A: Ukrainian

Q: [question]

A:

C.2.2 CoT, CoT-SC prompt

Question: Which film was released earlier, Kistimaat or I'M Taraneh, 15?

A: First, film Kistimaat was released on 6 October 2014. Second, I'M Taraneh, 15 was released on 13 November 2002. The answer is I'M Taraneh, 15.

Question: What is the date of death of the composer of film Baalaraajana Kathe?

A: First, the composer of film Baalaraajana Kathe is S. Hanumantha Rao. Second, S. Hanumantha Rao died on 27 May 1980. The answer is 27 May 1980.

Question: Who is the spouse of the director of film Alive (1993 Film)?

A: First, the director of film Alive is Frank Marshall. Second, Frank Marshall's spouse is Kathleen Kennedy. The answer is Kathleen Kennedy.

Question: Who lived longer, Edward Frederick Sanderson or Forrest Towns?

A: First, Edward Frederick Sanderson died at age 81. Second, Forrest Towns died at age 77. The answer is Edward Frederick Sanderson.

Question: Which country the director of film Battle Circus (Film) is from?

A: First, the director of film Battle Circus (Film) is Richard Brooks. Second, Richard Brooks was American. The answer is American.

Question: Which country the director of film In The Fog is from?

A: First, the director of film In The Fog is Sergei Loznitsa. Second, Sergei Loznitsa is Ukrainian. The answer is Ukrainian.

Question: [question]

A:

C.2.3 Verifying Question Generation prompt

Write a question that validates the reason for an overall question.

Overall Question: What is the date of death of the composer of film Baalaraajana Kathe?

Reason: First, the composer of film Baalaraajana Kathe is S. Hanumantha Rao.

Question: Who is the composer of film Baalaraajana Kathe?

Overall Question: Who lived longer, Edward Frederick Sanderson or Forrest Towns?

Reason: First, Edward Frederick Sanderson died at age 81.

Question: How long did Edward Frederick Sanderson live for?

Overall Question: [original question]

Reason: [rationale sentence]

Question:

C.2.4 Verifying Answer Generation (Rationale Editing) prompt

The film was released in 1984 by Essex Films. Kistimaat is a 2014 Bangladeshi action film directed by Ashiqur Rahman and produced by Tiger Media Limited and The Abhi Pictures. I'm Taraneh, 15 is a 2002 Iranian film directed by Rasul Sadrameli. The film was released on May 4, 2001.
Question: When was the film Kistimaat released?
Answer: The film Kistimaat was released in 2014.

Dwaram Venkataswami Naidu and also a lyricist. The film has musical score by S. Hanumantha Rao. Rao died 27 May 1980. Rao married Raja Mani with whom he had three daughters and one son.
Question: Who is the composer of film Baalaraajana Kathe?
Answer: The composer of film Baalaraajana Kathe is S. Hanumantha Rao.

Adib Kheir was a leading Syrian nationalist of the 1920s. Filmed on location in the Purcell Mountains in British Columbia, the film was directed by Frank Marshall, written by John Patrick Shanley, and narrated by John Malkovich. Frank Wilton Marshall(born September 13, 1946) is an American film producer and director, often working in collaboration with his wife, Kathleen Kennedy. He received the Irving G. Thalberg award from the Academy of Motion Picture Arts and Sciences in 2018.
Question: Who is the director of film Alive (1993 Film)?
Answer: The director of film Alive is Frank Marshall.

[context]
Question: [verifying question]
Answer:

C.3 Fever

C.3.1 Few-shot prompt

Determine if there is Observation that SUPPORTS or REFUTES a Claim, or if there is NOT ENOUGH INFO.

Claim: Reg Watson is a current television producer.
A: REFUTES

Claim: The Gadsden flag was named by Christopher Gadsden.
A: NOT ENOUGH INFO

Claim: Black Mirror is about society.
A: SUPPORTS

Claim: [question]
A:

C.3.2 CoT, CoT-SC prompt

Determine if there is Observation that SUPPORTS or REFUTES a Claim, or if there is NOT ENOUGH INFO.

Claim: Reg Watson is a current television producer.
A: First, Reginald James Watson AM was an Australian television producer and screenwriter. Second, Reginald James Watson AM died on 8 October 2019. The answer is REFUTES.

Claim: The Gadsden flag was named by Christopher Gadsden.
A: First, The Gadsden flag is named after politician Christopher Gadsden. Second, there is no information on who named the Gadsden flag. The answer is NOT ENOUGH INFO.

Claim: Black Mirror is about society.
A: First, Black Mirror is a British anthology television series. Second, The series uses technology to comment on contemporary social issues. The answer is SUPPORTS.

Claim: [question]
A:

C.3.3 Verifying Question Generation prompt

Write a question that validates the reason for a claim.

Claim: Reg Watson is a current television producer.
Reason: Reginald James Watson AM was an Australian television producer and screenwriter.
Question: What is Reg Watson's occupation?

Claim: The Gadsden flag was named by Christopher Gadsden.
Reason: there is no information on who named the Gadsden flag.
Question: Who named the Gadsden flag?

Claim: [question]
Reason: [rationale sentence]
Question:

Human Evaluation for Reasoning Chains

A Large Language Model is trying to answer a question, and it generated the following two reasoning chains before answering, which reasoning chain is more factually consistency in your opinion? Do you think that reasoning chain will lead to better answer predictions? (Remember, you can use google search to look it up! Please copy and paste the queries you searched for in question 3.)

Question: Roy Shepherd was considered a faculty member of what combination of colleges/universities?

Reasoning Chain 1: First, Roy Shepherd was considered a faculty member of the University of California, Berkeley, and Stanford University. Second, the combination of colleges/universities is the University of California, Berkeley and Stanford University.

Reasoning Chain 2: First, Roy Shepherd is a faculty member of a college/university with a combination of professional expertise and pedagogical excellence. Second, Roy Shepherd was a faculty member of the University of Notre Dame in Indiana.

(not shared) Switch account

* Required

1. Which reasoning chain is more factually consistent in your opinion? (you can use google search!)   *

○ Reason Chain 1

○ Reason Chain 2

○ tie

Figure 4: Example Screenshot of Human Evaluation User Interface.

C.3.4 Verifying Answer Generation (Rationale Editing) prompt

Reginald James Watson AM (27 August 1926 – 8 October 2019) was an Australian television producer and screenwriter. He was executive producer on Crossroads and created Australian media exports serials such as Prisoner, Neighbours, The Young Doctors and Sons and Daughters.

Question: What is Reg Watson's occupation?

Answer: Reg Watson was an Australian television producer and screenwriter

The flag is named after politician Christopher Gadsden (1724–1805), who designed it in 1775 during the American Revolution.

Question: Who named the Gadsden flag?

Answer: The Gadsden flag is named after Christopher Gadsden, but there is no information on who named it.

[context]

Question: [verifying question]

Answer:

D Human Study

To conduct the human study, we show the instructions in Fig. 4 to two human volunteers. The volunteers are NLP Ph.D. students who are proficient in English. The volunteers understand the use for the data collection and are in consensus. The reasoning chain 1 and 2 are CoTs generated by the CoT-SC baseline and the Verify-and-Edit shown in random order. On average, each volunteer took 1.25 hours to finish 50 samples.

E Qualitative Examples

In Table 5, 3 examples from the Adversarial HotpotQA datasets are shown in detail.

From the first sample, the LLM incorrectly states that the song is "based on .. Spider-Man." However, in the Google retrieved facts, it clearly states that it is based on "Ghost Rider". Therefore, the retrieved fact is able to help correct the detail in the rationale. Moreover, although the original rationale also covered the brand name "Marvel Comics", the generation goes on with the hero name as an answer, instead of the "brand" being asked. Feeding in again also corrects that logical mistake.

In the second example, the LLM makes up a plausible-sounding fact that "Tony Robinson has written seven children's books". There is also no indicator on the LLM's confidence level of this claim. Thus, if a user is unfamiliar with this knowledge, it could easily be mistaken as a true fact, which is highly risky. By introducing Google as an assistive tool, we retrieve the sentence "he has published 16 children's books." With this newly retrieved fact in mind, the LLM goes on generating the correct answer.

The third example is an interesting one. The original CoT already makes mistakes in the first sentence and goes on making continued mistakes in the second sentence as well. This is a type of common mistake in the dataset as well. On correcting them, the Verify-and-Edit framework is able to correct the first claim with the show "Chelsea Does". The second claim, however, is verified but irrelevant to the original question anymore. In this case, by feeding in both rationale sentences, the LLM is able to select the relevant fact as an answer, while disregarding the rest. This example shows that the CoT setup used by Verify-and-Edit is important as it allows for models to reason and abstract for a second time, instead of plainly replacing and correcting.

Original Question	Suicide's 1977 released album features a song based on what brand's comic character?
True Answer	Marvel Comics
CoT-SC	First, Suicide released their 1977 album titled Suicide: Alan Vega & Martin Rev.
	Second, the album features a song based on the Marvel Comics character Spider-Man.
	The answer is Spider-Man.
Verifying Questions	What album was released by Suicide in 1977?
	What comic character is the song on Suicide's 1977 album based on?
Retrieved Facts	Suicide is the debut studio album from the American rock band Suicide. It was released in 1977
	on Red Star Records and produced by Craig Leon and Marty Thau.
	"Ghost Rider" is a song by the protopunk band Suicide appearing on their debut album. The song
	is based on the Marvel Comics character. On Suicide, the noir-movie vibes of engine-revving tunes
	"Ghost Rider" (named after Vega's favorite comic book) and "Rocket U.S.A." feel scary.
Verify-and-edit	First, Suicide released their debut studio album, titled Suicide, in 1977.
	Second, The song on Suicide's 1977 album is based on the Marvel Comics character Ghost Rider.
	The answer is Marvel Comics.
Original Question	How many children's books has the writer of the sitcom Maid Marian and her Merry Men written?
True Answer	16
CoT-SC	First, the writer of the sitcom Maid Marian and her Merry Men is Tony Robinson.
	Second, Tony Robinson has written seven children's books.
	The answer is seven.
Verifying Questions	Who is the writer of the sitcom Maid Marian and her Merry Men?
	How many children's books has Tony Robinson written?
Retrieved Facts	Maid Marian and her Merry Men is a British children's television series created and written by Tony
	Robinson and directed by David Bell. It began in 1989 on BBC1 and ran for four series, with
	the last episode shown in 1994.
	He has published 16 children's books.
Verify-and-Edit	First, The writer of the sitcom Maid Marian and her Merry Men is Tony Robinson.
	Second, Tony Robinson has written 16 children's books.
	The answer is 16.
Original Question	What show other than Hello Ross did Chelsea Handler appear on in January of 2016
True Answer	Chelsea Does
CoT-SC	First, Chelsea Handler appeared on the show The Tonight Show Starring Jimmy Fallon in January
	of 2016.
	Second, The Tonight Show Starring Jimmy Fallon is a talk show hosted by Jimmy Fallon.
	The answer is The Tonight Show Starring Jimmy Fallon.
Verifying Questions	On which show did Chelsea Handler appear in January of 2016?
	What talk show is hosted by Jimmy Fallon?
Retrieved Facts	Chelsea Does is an American streaming television documentary series first released on Netflix on
	January 23, 2016. ... The episodes follow comedian Chelsea Handler as she explores different ...
	The Tonight Show Starring Jimmy Fallon is an American late-night talk show hosted by Jimmy
	Fallon that airs on NBC.
Verify-and-Edit	First, Chelsea Handler appeared on the show Chelsea Does in January of 2016.
	Second, The Tonight Show Starring Jimmy Fallon is a talk show hosted by Jimmy Fallon.
	The answer is Chelsea Does.

Table 5: Examples from AdvHotpotQA, facts are retrieved with Google.